1975

The Denaturation And Renaturation Of Deoxyribonucleic Acid Molecules

Lap-chung Michael Leung

THE DENATURATION AND RENATURATION OF DEOXYRIBONUCLEIC ACID MOLECULES

by

Michael Lap-Chung Leung

Department of Physics

Submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Faculty of Graduate Studies

The University of Western Ontario

London, Ontario

February, 1975

Dedicated to my parents, who encouraged me in this endeavour and in many others which I did less well....

ABSTRACT

Following the model first proposed by Zimm and others (1960), equilibrium properties of heterogeneous DNA molecules (i.e. those containing both AT and GC groups) near the melting temperature T_m are investigated using the grand partition function. Unlike previous authors who have resorted to numerical solution for the melting curves, we have succeeded in solving our model exactly and analytically. The algebraic expressions we obtained enable us to understand and correlate many observed equilibrium phenomena. The following quantities have been investigated.

(1) The melting temperature T_m. It is shown how the Marmur-Doty empirical relation between T_m and the concentration of GC content X_{GC} can be derived. The Marmur-Doty relation is a good approximation to the exact algebraic relation resulting from the Zimm model. A new convenient method of determining X_{GC} is proposed.

(2) Transition width W. Expressions for W in terms of X_{GC} and the parameter σ characterizing cooperativity among the base pairs have been obtained.

(3) Partial melting curves θ_1 and θ_2. These are the curves describing the percentage of "melted" AT-bonds θ_1 and the percentage of "melted" GC bonds θ_2 as a function of temperature. The approach of Felsenfeld and others in plotting θ_1 and θ_2 is examined. New insights have been obtained from our analytic expressions with relation to the base clustering question (i.e. referring to sections of DNA having distinctively different GC contents).

(4) Mean length h of a helical region and correlation length γ.

Here h is defined to be the average length of base pairs still in the helical form. The correlation length γ is the average distance between initiation points of consecutive helical regions. These quantities are believed to be observable and are related to viscosity measurements. Our theory indicates that the quantities h and γ are more sensitive to the cooperativity parameter σ than the transition width W is.

Dedicated to my parents, who encouraged me in this endeavour and in many others which I did less well....

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Prof. B.Y. Tong for introducing this problem and for his constant guidance and instructive advice throughout this work.

The helpful discussions with and continual encouragement from Dr. F.C. Choo are also well appreciated.

Thanks are also due to Prof. F.Y. Wu and Dr. M.F. Chiang for making a preprint of their work available and for their concern in this study.

I would also like to thank Prof. D.F. Gallaher and Prof. P.W. Whippey for their valuable suggestions throughout the preparation of the thesis.

Thanks are also extended to Prof. B.H. Zimm, Prof. E.W. Montroll, Prof. D.M. Crothers, Prof. M. Mendel and Prof. R.B. Inman for their interests and many valuable criticisms on this work.

This work is supported in part by a grant from the National Research Council of Canada to Prof. B.Y. Tong. This grant and the award of a National Research Council of Canada Postgraduate Scholarship are gratefully acknowledged.

I also wish to thank Miss M. Chan for typing the final drafts of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Traditionally, scientists have been reluctant to apply the quantitative methods of physics and chemistry to biological systems. This is because in most instances these are too complex to be examined from such fundamental standpoints. The mathematical analysis simply becomes too involved. Further, until recently, most biological data were of such a descriptive nature that they were not suitable for quantitative analysis. This situation has changed considerably during the last ten years or so. With the introduction of modern precision electronic equipment and recent advances in experimental technology, quantitative biology has now become possible. It is time to introduce developed and polished physical theories such as statistical mechanics to the study of biological systems (Bresler, 1970).

The DNA-solvent system, to be described in detail in the following chapters, is a relatively simple one from the biological standpoint. It offers an excellent example for the application of such ideas. Because of its great significance in genetics and molecular biology, a large amount of study has been undertaken on this system during the past two decades. In this thesis we shall concern ourselves with only one important aspect, the study of the breaking and reforming of the two component strands of the DNA double helix. This process is known as denaturation and renaturation*. We shall see that the process of

* Appendix V contains a list of those biological terms which may be unfamiliar to the reader with a physical science background.

denaturation and renaturation (also known as melting or helix-coil transition) and other closely related phenomena can be understood in terms of statistical mechanics. Information about the denaturation and renaturation process can be obtained from studies of the melting curve (loosely speaking, this is a curve showing the degree of separation between the two component strands of the DNA molecule as a function of temperature), the partial melting curves, mean lengths of helical regions and mean lengths of helix-plus-coil regions, and stacking energy between base pairs. Denaturation has been studied for the past twenty years. Most of these studies are concentrated on the determination of the melting curves. Even in this aspect of the denaturation process alone, systematic experimental studies are not sufficient. As most experimental investigations use DNA denaturation as a means towards the study of various other properties of the DNA molecule, it is not difficult to understand why the denaturation problem lacks unification when viewed as a whole. Experimental studies are hard to compare with one another because of the wide range of types of DNA and solvents used. Apart from the direct study on DNA melting, data and theories on other above mentioned areas connected with this process are very scarce.

There is yet no theory of denaturation and renaturation which, based upon a consistent set of parameters, satisfactorily serves the dual purposes of explanation and prediction. Because of the mathematical complexity, most calculations have been carried out using the digital computer. One of the inherent disadvantages of the numerical approach lies in the purely numerical results obtained. It is always preferable to be able to derive analytic results before one actually proceeds to

obtain numbers from them. Analytic results enhance one's physical insight, from which important inferences are developed.

The search for more accurate theoretical descriptions of the denaturation and renaturation process remains compelling, particularly in view of the following incentives. The arrangement of base pairs in the DNA molecule has a direct bearing on the genetic code. Studies to uncover the DNA base sequence have always been one of the top priorities in the fields of genetics and molecular biology. It is known that DNA molecules with different sequences have different melting profiles upon denaturation and renaturation. Therefore it is natural to believe that information on the base sequence may be obtained by studying the details of the melting process. In addition a clear picture of the equilibrium aspects of the melting process is helpful and may be essential in understanding the kinetics of DNA unwinding and replication. A direct application to other studies is also apparent. For example, the guanine-cytosine content X_{GC} of the DNA molecule occupies a unique position in bacterial taxonomy (Seidler and Mendel, 1971; Mendel, 1969), and this quantity can be obtained from melting curve studies (Marmur and Doty, 1962).

It is our purpose here to construct a unified theoretical model for the melting process and related phenomena. We do not treat these as separate entities. We have derived analytic results and expressions for the melting curves, partial melting curves, mean length of helical and helix-plus-coil regions, the transition width and other thermodynamical quantities. Based on a consideration of microscopic processes we derive a simple expression relating the guanine-cytosine content X_{GC} to the

melting temperature T_m. We have found that this expression is compatible with the well known empirical relation first presented by Marmur and Doty (1962). In addition, consideration of one of our theoretical expressions leads us to propose a new experimental method for the determination of the guanine-cytosine content X_{GC}. This seems simpler than the conventional melting temperature experiment. The stacking interaction among base pairs is an area that is still uncertain both experimentally and theoretically. From the present study, we are able to say that variations in this quantity U have the greatest effects on the mean length of a helical region, h. They also effect the transition width W but have little effect on the melting temperature. Thus the estimation of the stacking energy via transition width measurements would be inferior to the method of determining the stacking energy through measurements of the mean length of a helical region. The theoretical models proposed by various authors disagree in their prediction for the temperature dependence of the base pair bond free energy (Montroll et al, 1966,1968; Vedenov et al, 1967, 1972; Lehman, 1967, Lehman and McTague, 1968). We shall suggest an experimental method of settling this argument via a careful analysis of the melting curves.

It is our hope that this work will lead to re-examination and revision of experiments using modern techniques, and also to new experiments.

CHAPTER II

EXISTING EXPERIMENTAL INFORMATION

We shall begin with a brief description of the biochemistry of the DNA molecule (Steiner and Beers, 1961; Watson and Crick, 1953; Watson, 1965). This macromolecule belongs to a class of important biopolymers known as the nucleic acids. It is a giant compound of molecular weight around 10^9. With reference to Fig.2.1, it is a complex of two chains, each being a linear polymer of nucleotides. A nucleotide in general consists of a sugar, a phosphate radical and a nitrogenous base. In the case of DNA, the sugar is deoxyribose and the nitrogenous base must be one of the four types: the purines adenine (A) and guanine (G), or the pyrimidines thymine (T) and cytosine (C). DNA molecules taken from different biological organisms differ only in two aspects: the extent of polymerization, and the sequence of its bases. The polynucleotide chains form two anti-parallel α-helices, mutually twisting around each other in a right-handed manner. The stability of this double helical structure is maintained by interstrand hydrogen bonds formed between one base group on one chain and another base group on the other, and reinforced by the cooperative stacking interaction among these base-pairs. This stacking interaction is a measure of how strongly a base pair interacts with its neighbours. The stronger the interaction, the more the tendency for a base pair to assume the state (broken or intact) as its neighbours. The base groups adenine (A) and thymine (T) as well as the groups guanine (G) and cytosine (C) are said to be complementary, in the sense that an

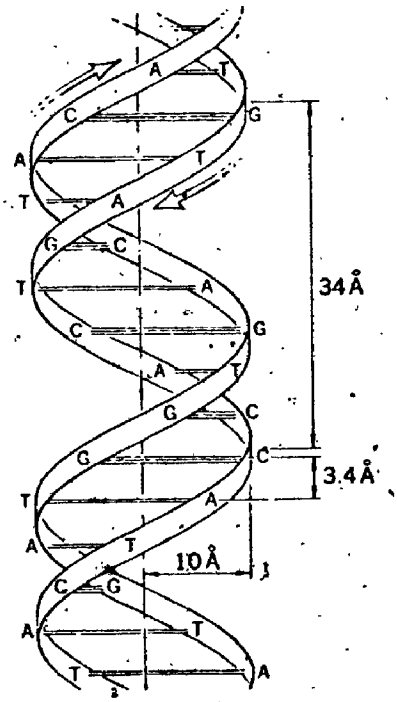


FIGURE 2.1 SCHEMATIC DIAGRAM OF A DNA MOLECULE.

A group is always bonded to a T group and a C group always to a G group. This is because hydrogen bonds form much more readily between complementary bases than non-complementary ones. This phenomenon is known as matching. In fact, mis-matching seldom occurs in natural DNA's. The GC pair is stronger, with three hydrogen bonds. The AT pair has only two. A typical DNA molecule may have as many as 10^5 base pairs. The diameter of the double helix is about 20Å and the pitch 34Å. The axial distance between neighbouring base pairs is 3.4Å while the distance if measured along the chain is 6.8Å. There are ten base pairs in every complete turn (see Fig. 2.1).

The DNA molecule can also exist in the denatured or coil state under appropriate conditions. Here, the two strands are physically separated and the configuration of each strand is similar to a random coil. Compared to the helical state, the coil state is very flexible. Now that the strand is no longer in the ordered helical state, the notion of a axial separation between base pairs is no longer meaningful. The distance between neighbouring bases along the chain remains to be 6.8Å.

The process of denaturation or melting involves the separation of the two complementary strands. This results from the rupture of the two forces mentioned above, namely, the hydrogen bonds holding complementary bases together to form base pairs, and the cooperative stacking interaction between the neighbouring base pairs. In the literature, the process is also known as the "helix-coil transition", since it involves the transition from the natural double helical state to the state of two random coils of polynucleotide chains. It may be brought about by the application of heat, change in pH value of the solvent or the introduction of certain chemicals.

The reverse process is known as renaturation. Renaturation can be brought about by inverting the experimental conditions. Given enough time, the complementary strands will reassociate to form the original double helical structure. The renaturation process is much faster if the previous denaturation has not been carried out to the extreme, so that the two strands are still physically connected by a few base pairs here and there. This transition process involves a competition between two opposing influences. Energetically, the helical state is more favourable because of its hydrogen bonds and positive cooperativity. By positive cooperativity, we mean that the base pairs tend to interact in such a way that their states of being broken or intact are alike, as opposed to negative cooperativity where the interacting units tend to align themselves in different states. However, because of its high flexibility, the coil state is favoured from the condition of maximum entropy. The transition is therefore reversible from the energetically favourable form of the double helix to the high entropy form of the random coil. The double helix resembles a highly ordered crystal while the random coil corresponds to a disordered liquid. It must be remembered that the sequence or order of bases in the polymer remains unaltered in either form, as only the weaker intramolecular forces are destroyed upon melting, while the strong covalent bonds in the sugar phosphate backbone remain unaffected.

A variety of experiments may be performed in connection with the helix-coil transition from which a number of characteristic quantities can be determined. These are summarised in the following:

(a) The melting curve.

The transition can be traced by observing changes in optical

activity, viscosity or sedimentation constant of the DNA-solvent system. It can also be followed by experiments in electron microscopy, microcalorimetry, chromatography and spectrophotometry. A melting curve is generally plotted and used to characterize the transition. Briefly, the melting curve gives the fraction or percentage of disrupted base pairs, $\theta(T)$, as a function of temperature T.

Perhaps the most commonly employed technique is the measurement of DNA absorbance in the near ultraviolet region of the spectrum centred around 260mμ. An increase in absorbance has been reported as the DNA denaturates, a phenomenon known as hypochromism. By recording this change, it is possible to construct the variation of $\theta(T)$ with T. If D_{min} (D_{max}) is minimum (maximum) values of the absorbance D, corresponding to all the base pairs being intact (disrupted), then

$$\theta(T) = (D - D_{min})/(D_{max} - D_{min}) \qquad (2.1)$$

Fig. 2.2 gives a typical "melting curve" so obtained.

During the transition, the double-stranded helix splits into two complementary strands. One therefore expects that those physical quantities which are attributes of the macromolecule's structural characteristics would likewise undergo some "transition". This is indeed the case. For example, the characteristic viscosity and sedimentation constant have been reported to undergo some considerable changes during the transition (Shaguili et al, 1969 and 1971). These

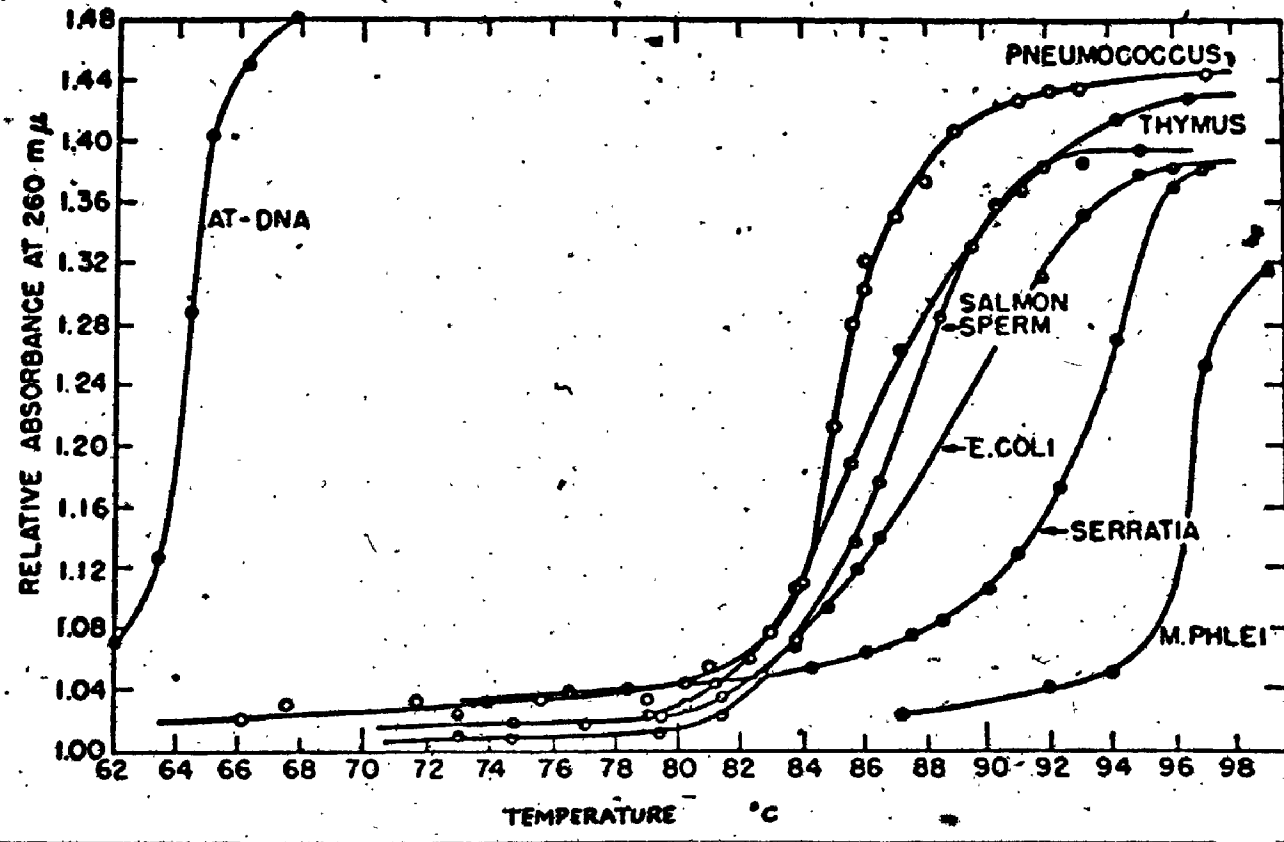


FIGURE 2.2 MELTING CURVES OF VARIOUS DNA MOLECULES IN SSC (FROM MARMUR AND DOTY)

changes can be also used to characterize the transition itself. In addition to these methods, the helix-coil transition has also been studied by the means of micro-calorimetry and electron microscopy experiments. It should be pointed out that this helix-coil transition is only accompanied by a sharp change in the physical quantities concerned and is quite different from what one generally knows as a phase transition as discussed in statistical mechanics or thermodynamics.

(b) The transition temperature or melting point, T_m.

Besides $\theta(T)$, another parameter that is useful to charaterize a melting curve is the "melting point", T_m. Because the transition is in fact not a true phase transition, from the thermodynamical sense, the definition of T_m is not unique and is still disputed. There are two major conventions adopted:

(i) T_m is the temperature at which exactly half the total number of base pairs have been disrupted and the other half remains intact (Marmur and Doty, 1962), i.e. $\theta(T_m)=0.5$.

(ii) T_m is the temperature at which the melting curve has an inflection point. This occurs at the point of maximum slope (Vedenov et al, 1972), i.e. when $\theta''(T_m)=0$.

These two definitions converge for the case of a simple symmetrical melting curve. However for more complicated curves, they give different T_m's. We have found the first definition easier to work with. This is so mainly because of two reasons: firstly, a complicated melting curve (e.g. see Boublik et al,

1965) like that of DNA from calf thymus may exhibit a number of inflection points rendering the second convention ambiguous; secondly, the sharpness of the transition often presents some difficulty in pin-pointing the inflection point. We shall therefore adopt the first alternative for the definition of T_m.

The GC pair is thermally more stable than the AT pair, partly because of its extra hydrogen bond. For this reason the melting temperature of the GC homopolynucleotide is found to be generally higher than that of the AT homopolynucleotide. This statement needs some qualification, since the melting temperature is also sensitively solvent dependent. In some solvents the difference in T_m can be as much as 40°C, while in others this difference may be zero. In general, for a given solvent, the value of T_m depends on the base composition of the DNA. However, it is observed that while this is so, the exact arrangement of the base sequence has little or no appreciable effect on T_m. Under similar conditions, DNA molecules having similar AT:GC ratio but different sequence arrangements melt at approximately the same temperature. Chemical features of the solvent also influence the value of T_m. This includes the pH of the solution as well as the concentration of ions. The dependence of T_m on percentage G-C content (X_{GC}) and the cation concentration has been found to be approximately linear, an observation made empirically (Marmur and Doty, 1962; Owen et al, 1969). The dependence of T_m on the pH of the solution is a non-linear one, T_m being

maximal in neutral pH, but dropping significantly at pH values below 5 or above 9.

Due to the diverging interest and purposes of different investigators, few thermodynamical properties of DNA are measured under identical conditions. This has led to complication and difficulty in making a systematic analysis involving comparison and interpretation of parameters. Nevertheless, a standard solvent condition comparable to that inside a cell has been agreed for melting curve measurements (Marmur and Doty, 1962). It is taken to be an aqueous solution containing 0.15M sodium chloride and 0.015M sodium citrate (abbreviated as SSC for Standard Saline Concentration). In SSC, the experimental values of T_m for various DNA's fall within the range from 342K to 384K. To a first approximation, the following empirical formula holds at neutral pH. (Vedenov, 1972):

$$T_m = 449.0 - (1.60 + X)(36.0 - 7.04 \log|Na^+|) \qquad (2.2)$$

where T_m is expressed in K, X is the fractional AT content and $|Na^+|$ is the molar concentration of sodium ions in the solution. (see Fig. 2.3 and Fig. 2.4) We shall see that in SSC conditions, our result gives Eq. (2.2) as a limiting case.

(c) The transition width, W.

Another parameter that characterizes a melting curve is the transition width, W. Again, as in the case of T_m, there is a disagreement over the definition of W:

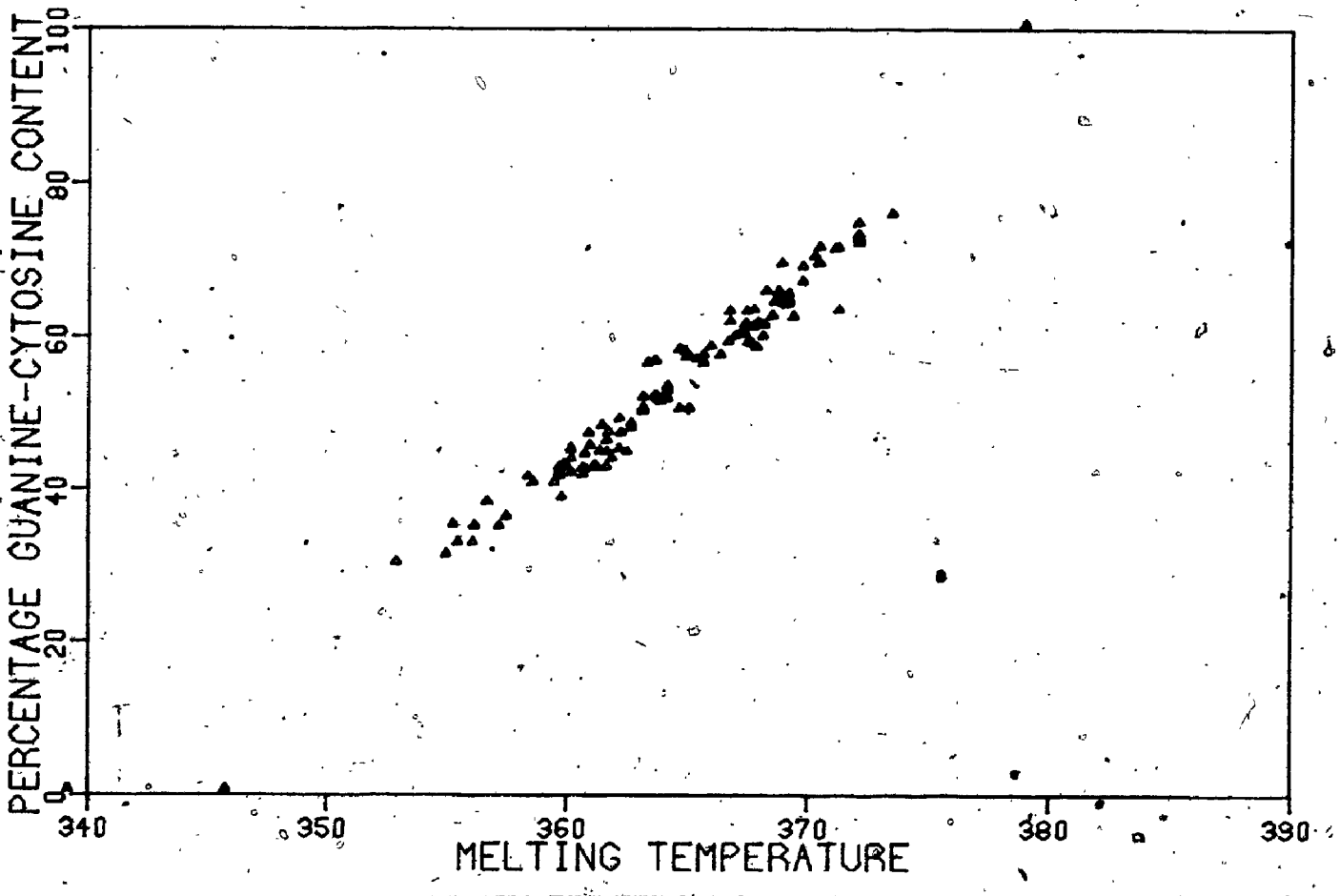


FIGURE 2.3 VARIATION OF MELTING TEMPERATURE WITH PERCENTAGE GC CONTENT
△ DATA OF DE LEY.

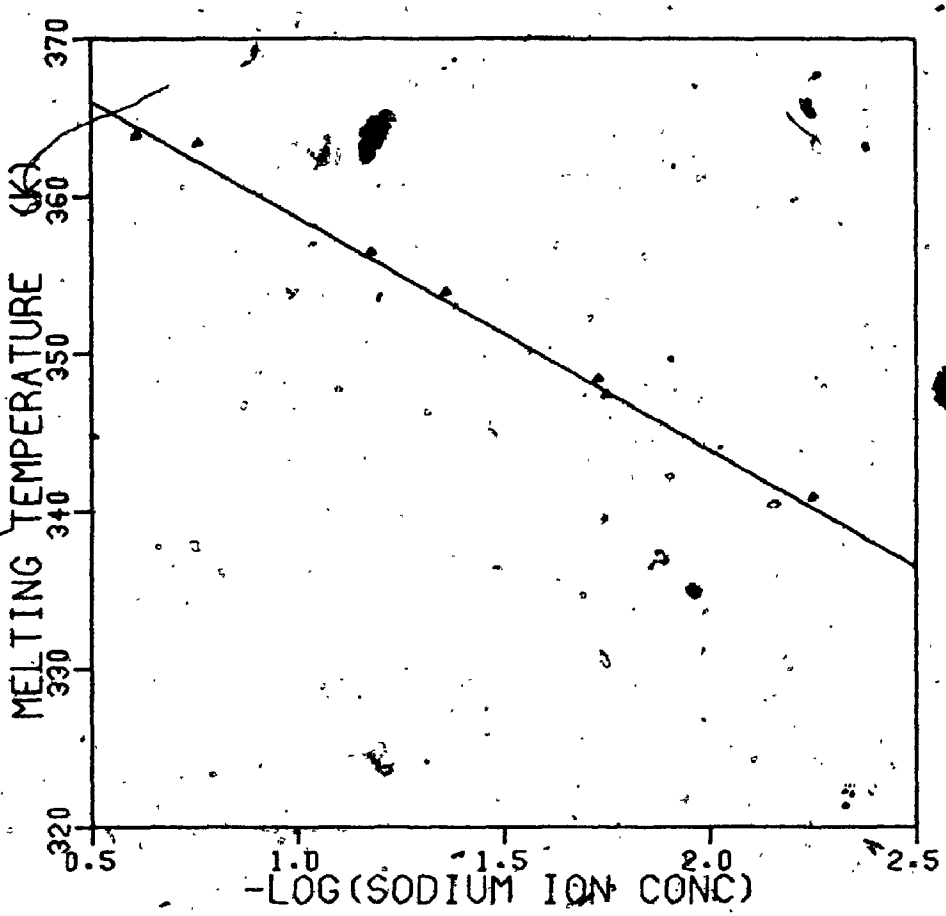


FIGURE 2.4 VARIATION OF MELTING TEMPERATURE WITH SODIUM ION CONCENTRATION
△ DATA OF OWEN ET AL FOR E. COLI
SOLID LINE IS EQ. (2.2) FOR X=0.5

(i) W is the difference between the temperature coordinates of the points at which the tangent to the melting curve drawn at $\theta(T)=\frac{1}{2}$ intersects the ordinates $\theta(T)=0$ and 1. Mathematically,

$$W = 1/\left|\frac{d\theta}{dT}\right| \quad \text{at } \theta = 0.5 \qquad (2.3)$$

(ii) Similar to definition (i) above, but the tangent is drawn at the point of inflection, thus giving

$$W = 1/\left|\frac{d\theta}{dT}\right|_{max} \qquad (2.4)$$

(iii) W is the temperature interval between two preassigned values of θ, say $\theta = 0.3$ and $\theta = 0.7$ (Reiss et al, 1966)

$$W = T(\theta=0.7) - T(\theta=0.3) \qquad (2.5)$$

(iv)

$$W = \left(T\left|\frac{d\theta}{dT}\right|\right)^{-1}_{T=T_m} \qquad (2.6)$$

The individual merits of each definition need no further elaboration. The criteria for the choice depends partly on how one previouly defined T_m and partly on what kind of studies one is interested in. All of these definitions have a single common aspect, with all of them embodying some measure of the "sharpness" of the transition. In what follows we shall adopt definition (i).

It is observed that W is broadest when the AT and GC base pairs are present in equal amounts and it is narrowest for the case of homopolynucleotides, which consist of only one kind of

base pairs, either GC or AT (Fig. 2.5). In SSC, W ranges from less than 1° to about $10^{\circ}K$. In general, two DNA molecules with different base sequence arrangements will have different transition widths, even if they do have comparable percentage G-C content. Another factor that influences the width is the degree of cooperativity σ. The degree of cooperativity σ is a numerical parameter which is a measure of how strongly a base pair interacts with its nearest neighbours. The stronger the interaction the greater the tendency for a base pair to assume the same state (broken or intact) as its neighbouring base pairs, and correspondingly the narrower the transition width W. A mathematical definition of σ will be found in Chapter III.

(d) Partial Melting Curves.

From studies of data obtained from spectrometric experiments, it is possible to determine the melting curves for the two types of base pairs separately (Felsenfeld and Sandeen, 1962). Consequently, the temperature dependence of the fraction of disrupted AT pairs, $\theta_1(T)$ (and of GC pairs, $\theta_2(T)$) may be displayed (Fig. 2.6). It is found that these two partial curves for θ_1 and θ_2 sandwich the overall melting curve θ for the whole DNA molecule. All three curves exhibit a similar general shape. It is a remarkable fact that the separation between th AT and GC melting curves in the temperature axis when $\theta=\frac{1}{2}$ is significantly narrower (about 5 degrees in the case of calf thymus DNA in low ionic strength) than the separation between melting curves of AT and GC homopolymers,

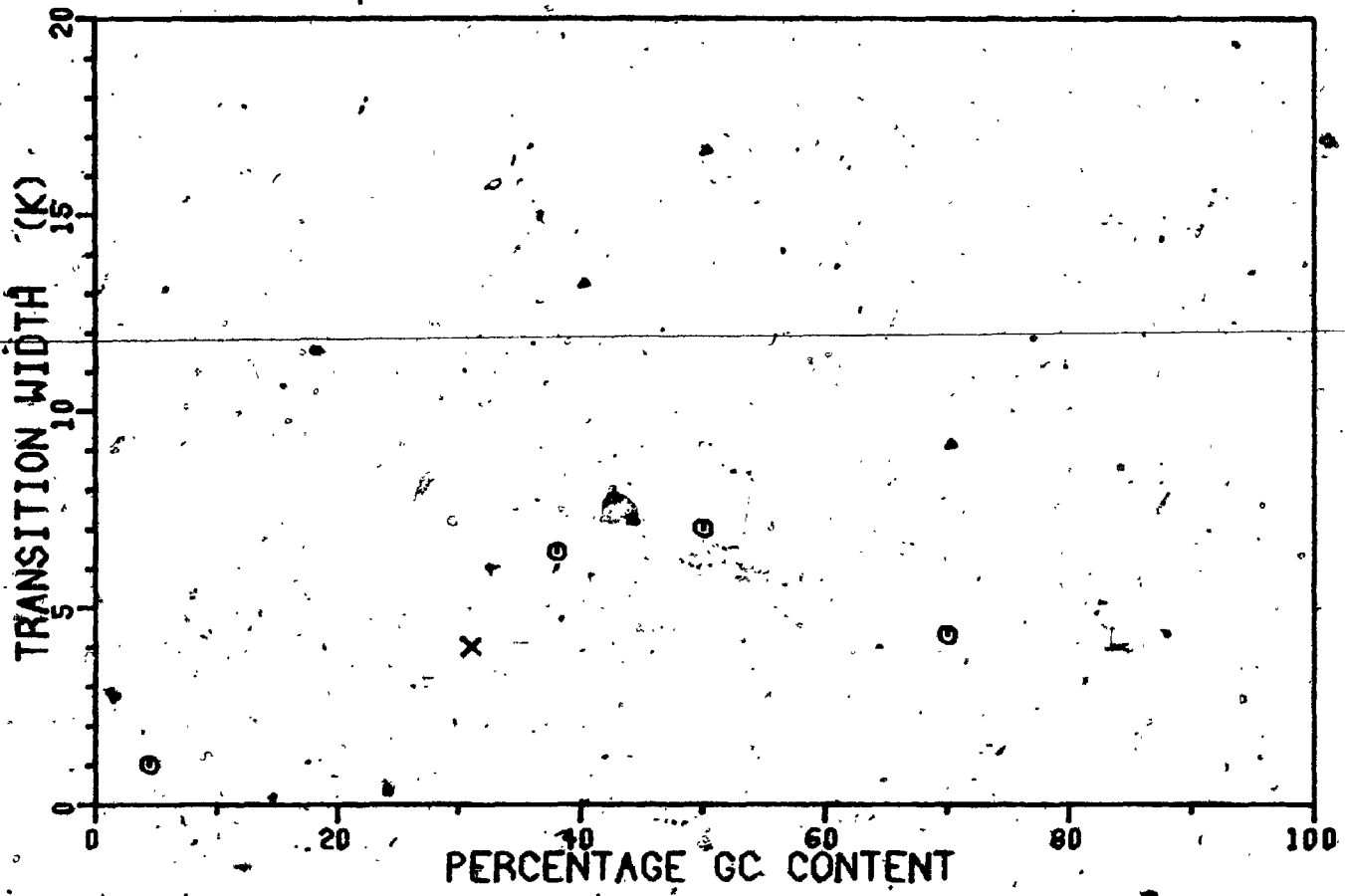


FIGURE 2.5 VARIATION OF TRANSITION WIDTH WITH PERCENTAGE GUANINE-CYTOSINE CONTENT
DATA OF DAVIDSON ET AL:
● USING 0.1M $NaClO_4$
△ USING $3X10^{-4}$M $NaClO_4$
× DATA OF CROTHERS ET AL. USING 0.018M NaCl.

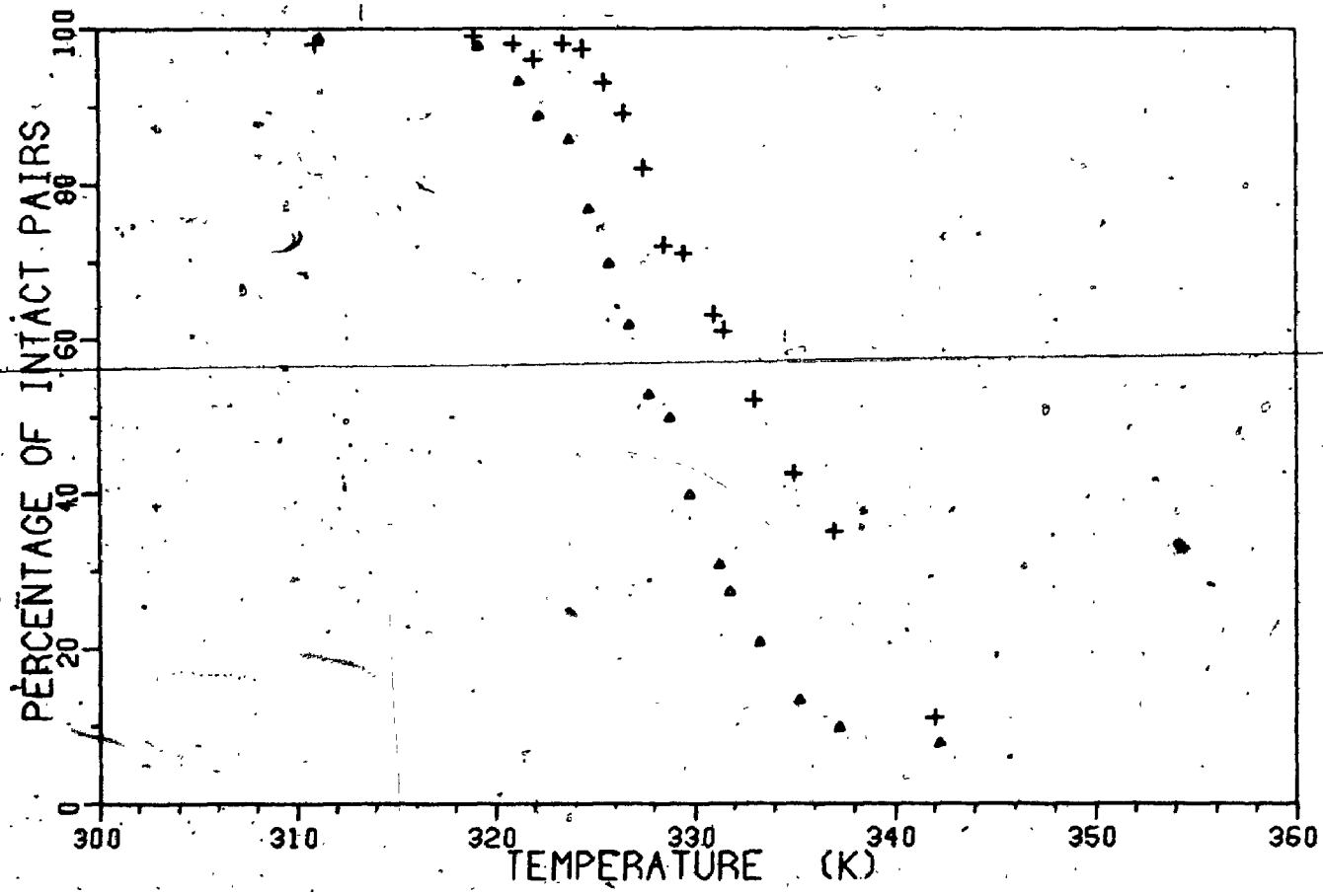


FIGURE 2.6 DATA OF FELSENFELD ET AL ON PARTIAL MELTING CURVES OF CALF THYMUS DNA
△ AT + GC

($41^{\circ}C$ in SSC). This observation shows that the denaturation behaviour of the AT and GC pairs is closely related, since the mean melting temperature of the AT is brought up and that of the GC is brought down, from their values in the homopolynucleotides. Another interesting observation is the non-linear relationship existing between θ_1 and θ_2 (Felsenfeld and Hirschman, 1965) (Fig. 2.7). This relationship indicates that the DNA does not melt uniformly, but that regions with higher AT content melt first.

(e) Mean length of a helical region, h.

During denaturation or renaturation, the DNA molecule is divided into consecutive helical and coil segemnts. This quantity h is defined to be the average length (measured in terms of number of base pairs) of such a helical segment. It is possible by parallel experiments using optical and viscosimetric techniques to determine the mean length of a helical region, h (Shagulii el al, 1969 and 1971) (Fig. 2.8). However, the results are not entirely satisfactory as other effects tend to obscure accurate data interpretation. Our results indicate that h is very sensitive to changes in the degree of cooperativity as well as in the DNA base sequence.

(f) Mean length γ of a helix-plus-coil region.

This quantity can also be meaured in a way similar to h. It is defined to be the mean number of links between consecutive initiation points of a helical region during the denaturation process. It is thus the average length of consecutive pairs of helix-plus-coil regions.

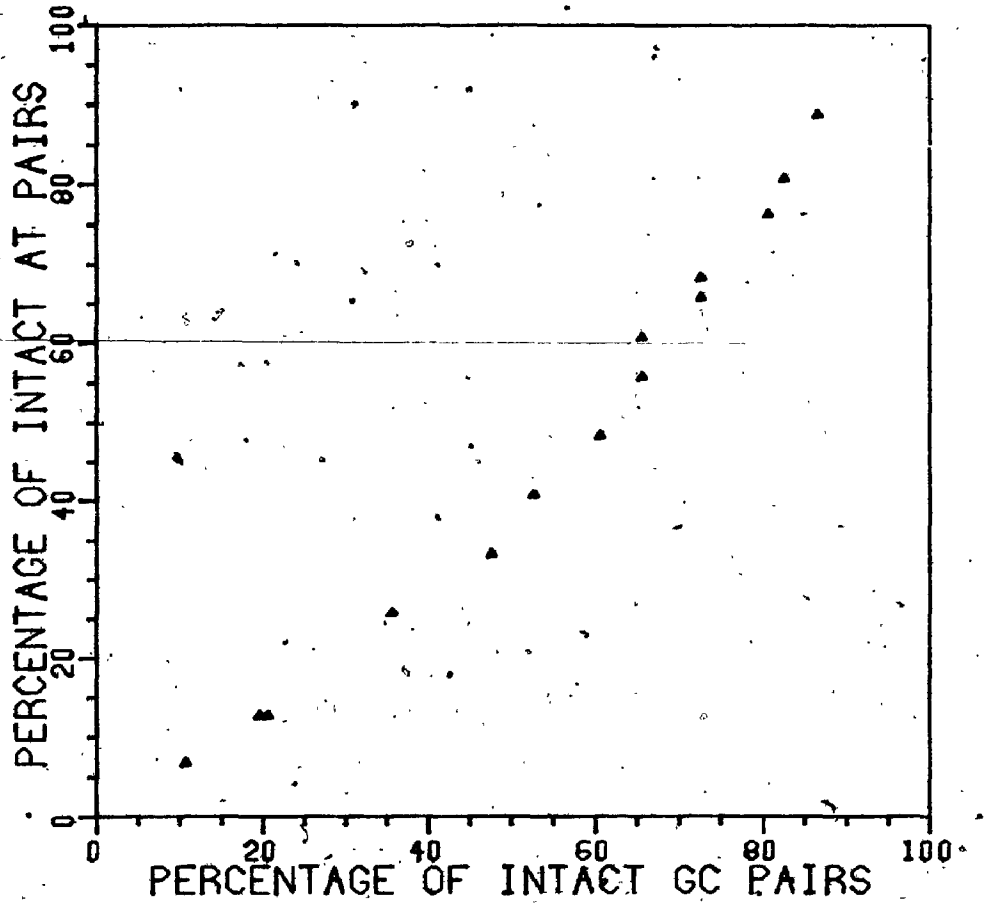


FIGURE 2.7 .INTACT AT AND GC BASE PAIRS

△ DATA OF HIRSCHMAN ET AL

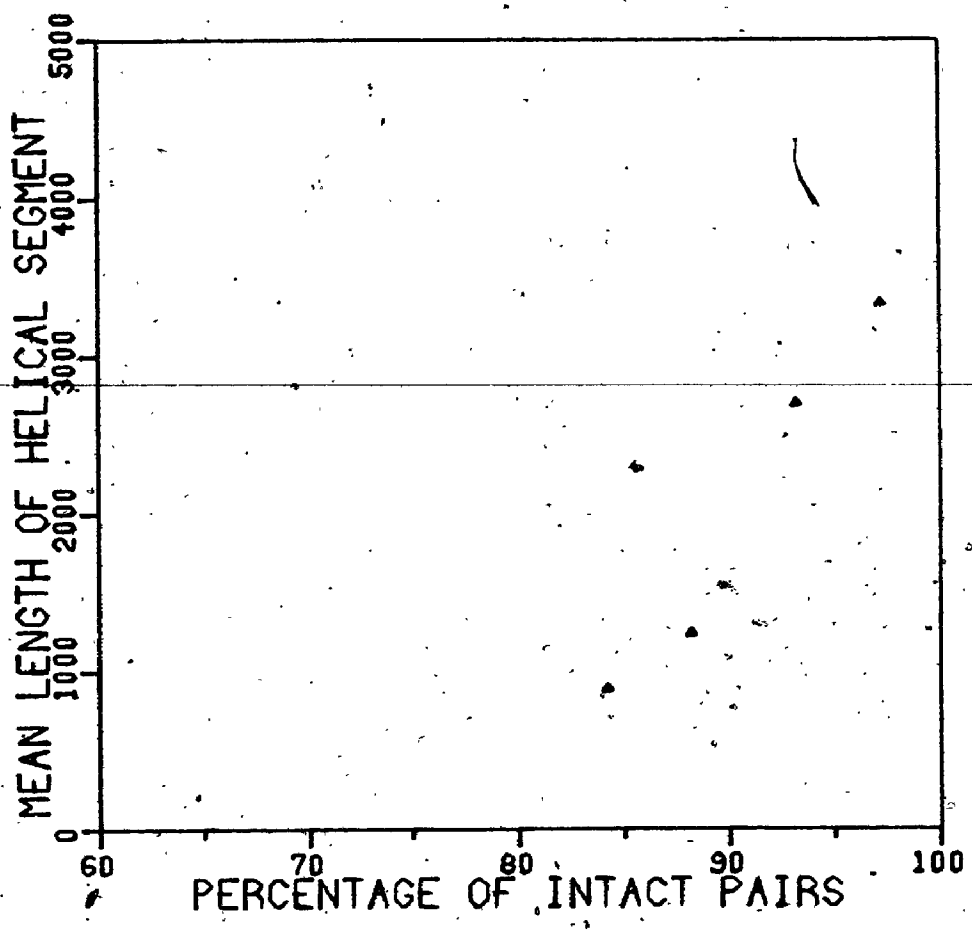


FIGURE 2.8 ▲ DATA OF SHAGULII ET AL FOR T2 PHAGE DNA.

The above discussion related to the seven parameters θ, T_m, W, θ_1, θ_2, h and γ represents a summary of the established experimental facts related to the helix-coil transition in the DNA molecule. It follows that any successful theory must be able to take all of them into account. Among them, the difference in thermal stability of AT and GC pairs, the variation of the transition width among different DNA species and the effects of the degree of cooperativity on parameters associated with the transition, etc., ought to be given appropriate weights in the consideration.

CHAPTER III

GENERAL FORMALISM OF THE MODIFIED ISING MODEL

The DNA molecule may be conveniently represented by a one-dimensional aperiodic binary substitutional alloy: binary since there are two types of unit cells, namely the AT and GC base pairs; one-dimensional because each cell has exactly two nearest neighbours; and substitutional since the sizes of AT and GC base pairs are almost identical (the length of an AT pair is 11.1 Å and that of a GC pair is 10.8 Å). The sequence of base pairs in natural DNA's has to be aperiodic because it is the carrier of the complete genetic information for the organism, loosely similar to the sequence of letters in a meaningful text being the carrier of the writer's message.

Before discussing the case of heterogeneous DNA, we must say a few words about homopolynucleotides, of which there are two types: that containing only the guanine-cytosine (GC) base pairs and that containing only the adenine-thymine (AT) base pairs. Because of the greater thermal stability of GC pairs, the GC homopolymer tends to melt at a higher temperature than the AT homopolymer. In SSC conditions, for example, the GC homopolymer melts at approximately 383 K while its AT counterpart melts at 342 K. The melting curves for these homopolymers are relatively sharp: the denaturation process is completed within a range of one degree. It therefore becomes interesting to investigate the melting behaviour of a DNA molecule containing both AT and GC base pairs and to compare the melting characteristics of the heterogeneous DNA with those of the homopolymers. It might also be interesting to study the influence of the base sequence and the degree of heterogeneity on the melting characteristics.

The relative amounts of the two base pair types present, as well as how they are arranged in the DNA molecule, appear to play an important role in determining the melting characteristics. This can be seen from a study of Fig. 3.1. Conversely, it is anticipated that a study of these melting characteristics will yield information about the percentage of GC content and/or the sequence of base pairs. The latter information on the sequencing is of particular importance to genetics. Unfortunately, up to the present moment, it has not yet been proved possible to completely identify the sequence of the 10^4 to 10^6 base pairs in a typical DNA molecule.

One approach from the theoretical standpoint is to consider the base sequence to be random. Lehman and McTague (1968) presented a calculation for random sequence DNA's with given GC concentration, X_{GC}. In their work, a base sequence was said to be random if the probability of a site being occupied by a particular base pair was independent of the base pairs occupying neighbouring sites. In our present calculation, a different approach has been adopted. We shall approximate the actual base sequence of a DNA with a given fixed concentration of GC base pairs, X_{GC}, by the statistically most probable sequence which is calculated as the ensemble average of all possible sequences. We note here that if the stacking interaction U is independent of the nature of the neighbouring base pairs, this approximation of the actual sequence is equivalent to the approximation by the averaged sequence of all random sequences with the same X_{GC}. Another possible approach would be to approximate the actual sequence by a realization of a random sequence generated via some computer routine (Fink and Crothers, 1968). This would not be so ideal because calculations using

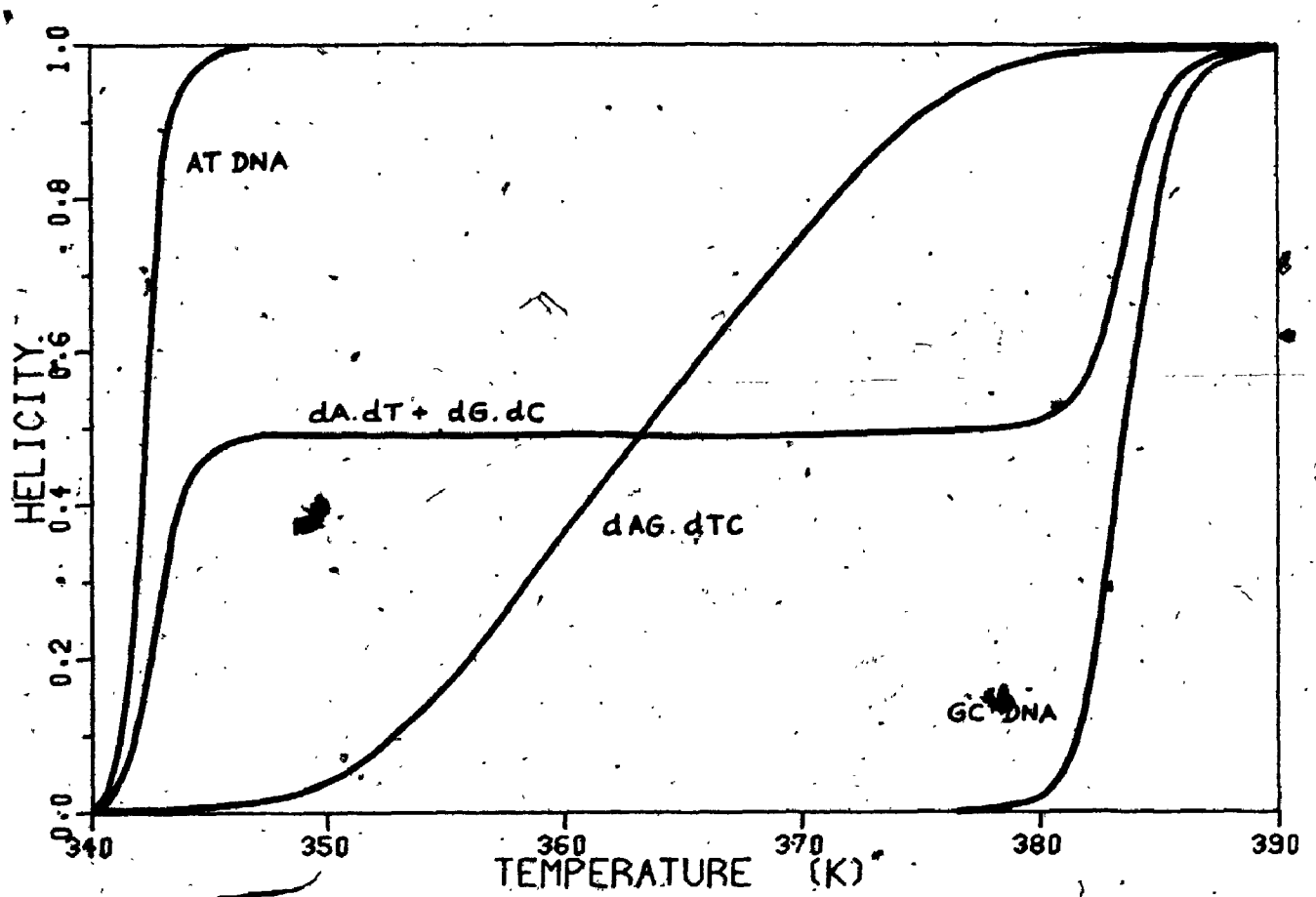


FIGURE 3.1 MELTING CURVES OF VARIOUS DNA MOLECULES.

different realizations of a random sequence would be found to give different results and therefore would imply an ambiguity in the final results. Furthermore, since the random sequence used would probably be different from the actual DNA sequence, one could never be sure of how good the approximation was.

We believe that the best test for an approach is the agreement with known experimental results. To this end we have a few areas at our disposal, as outlined in Chapter II. It is our goal to give estimates of such key parameters as the cooperativity factor σ, the magnitude of which is still not certain from experiments. Some authors advocate that different species of DNA have different σ values while others believe that a common σ for all species would be more appropriate. In addition, it is also our intention to develop a consistent and unified theoretical framework for the denaturation process and related phenomena.

To begin with, we shall consider a base pair to be intact when both the horizontal (hydrogen bonds) and vertical (stacking) interactions are present; otherwise the base pair is considered to be totally disrupted. It is assumed that there is no intermediate stage. The state of a base pair (intact or disrupted) will be designated by the symbol τ with $\tau = 1$ for an intact pair and $\tau = -1$ for a disrupted pair. We shall presently elaborate on the grand partition function method developed by Chiang and Wu (Chiang, Ph.D. thesis, 1968).

Let A and B denote the sets of link states (intact or disrupted) of the AT and GC links respectively, so that a given link in the sequence may assume any one of the following four possible states, namely: AT, intact or disrupted, GC intact or disrupted. Accounting only for nearest neighbour interactions, the Hamiltonian for the whole molecule consisting

of N links is of the Ising type and is given by

$$H = -\sum_{j=1}^{N} \{U(\tau_j, \tau_{j+1})\tau_j \tau_{j+1} + J(\tau_j)\tau_j\}, \qquad \tau = \pm 1, \tag{3.1}$$

where

$$U(\tau_j, \tau_{j+1}) = \begin{cases} \tilde{U}_1 & \text{if } \tau_j, \tau_{j+1} \in \{A\} \\ \tilde{U}_2 & \text{if } \tau_j, \tau_{j+1} \in \{B\} \\ \tilde{U}_3 & \text{otherwise,} \end{cases} \tag{3.2}$$

$$J(\tau_j) = \begin{cases} \tilde{J}_1 & \text{if } \tau_j \in \{A\} \\ \tilde{J}_2 & \text{if } \tau_j \in \{B\} \end{cases} \tag{3.3}$$

and {A} and {B} denote the set of base pair states belonging to the AT type and GC type respectively. The grand partition function Y' is

$$Y' = \sum_{n=0}^{N} \sum_{\{\tau_j\}} \exp(-\beta H) \exp\{\beta|\mu_{AT} n + \mu_{GC}(N-n)|\} \, \delta\left(\sum_{j=1}^{N} g(\tau_j) - n\right) \tag{3.4}$$

where μ_{AT}, μ_{GC} are the chemical potentials for AT and GC pairs,

n is the number of AT pairs in a given configuration,

$\{\tau_j\}$ denotes the set of all possible states for a given n,

$$g(\tau_j) = \begin{cases} 1 & \text{if } \tau_j \in \{A\} \\ 0 & \text{if } \tau_j \in \{B\} \end{cases}$$

and δ is the usual Kronecker delta function ensuring the fulfilment of the condition

$$n = \sum_{j=1}^{N} g(\tau_j) \tag{3.5}$$

The following form of the grand partition function will be found more useful:

$$Y = \exp(-\beta\mu_{GC}N)\, Y' = \sum_{n=0}^{N} \sum_{\{\tau_j\}} \exp(-\beta H)\exp\{\beta\mu \sum_{j=1}^{N} g(\tau_j)\} \qquad (3.6)$$

in which $\mu = \mu_{AT} - \mu_{GC}$.

Now define a 4×4 matrix Q with components

$$Q(\tau,\tau') = \exp\{\beta U(\tau,\tau')\tau\tau' + \tfrac{1}{2}\beta[J(\tau)\tau + J(\tau')\tau'] + \tfrac{1}{2}\beta\mu[g(\tau)+g(\tau')]\} \qquad (3.7)$$

Written explicitly,

$$Q = \begin{pmatrix} y_1 p_1 z & z/y_1 & y_3 p_1^{\frac{1}{2}} p_2^{\frac{1}{2}} z^{\frac{1}{2}} & p_1^{\frac{1}{2}} z^{\frac{1}{2}}/y_3 p_2^{\frac{1}{2}} \\ z/y_1 & y_1 z/p_1 & p_2^{\frac{1}{2}} z^{\frac{1}{2}}/y_3 p_1^{\frac{1}{2}} & y_3 z^{\frac{1}{2}}/p_1^{\frac{1}{2}} p_2^{\frac{1}{2}} \\ y_3 p_1^{\frac{1}{2}} p_2^{\frac{1}{2}} z^{\frac{1}{2}} & p_2^{\frac{1}{2}} z^{\frac{1}{2}}/y_3 p_1^{\frac{1}{2}} & y_2 p_2 & 1/y_2 \\ p_1^{\frac{1}{2}} z^{\frac{1}{2}}/y_3 p_2^{\frac{1}{2}} & y_3 z^{\frac{1}{2}}/p_1^{\frac{1}{2}} p_2^{\frac{1}{2}} & 1/y_2 & y_2/p_2 \end{pmatrix} \qquad (3.8)$$

in which the fugacity ratio z is defined as

$$z = \exp(\beta\mu)$$

and

$$y_i = \exp(\beta\tilde{U}_i) \equiv \exp(U_i) \qquad i = 1,2,3. \qquad (3.9)$$

$$p_\ell = \exp(\beta\tilde{J}_\ell) \equiv \exp(J_\ell) \qquad \ell = 1,2 .$$

Q is a symmetric matrix with positive entries. Using periodic boundary conditions Y can be identified with trace(Q^N), (Huang, 1963). For natural DNA's N is typically of the order 10^4 to 10^6. It will therefore

be reasonable to make the simplifying assumption that $N \to \infty$. If $\bar{\lambda}$ is the largest eigenvalue of Q, then we may make the following approximation:

$$\lim_{N\to\infty} (\ln Y)/N \simeq \ln\bar{\lambda} \qquad (3.10)$$

The problem therefore reduces to evaluate $\bar{\lambda}$ in terms of the known parameters. After tedious but straight forward algebraic manipulation, the secular equation for the matrix Q can be simplified into

$$A_4\lambda^4 + A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0 = 0 \qquad (3.11)$$

where $A_4 = 1$

$$A_3 = -2(c_1y_1z + c_2y_2)$$

$$A_2 = (y_1^2 - y_1^{-2})z^2 + 2c_+z(y_1y_2 - y_3^2) + 2c_-z(y_1y_2 - y_3^{-2}) + y_2 - y_2^{-2}$$

$$A_1 = -2\{c_2z^2[y_2(y_1^2 - y_1^{-2}) - y_1(y_3 + y_3^{-1})^2 + 2(y_1 + y_1^{-1})]$$
$$+ c_2z[y_1(y_2^2 - y_2^{-2}) - y_2(y_3 + y_3^{-1})^2 + 2(y_2 + y_2^{-1})]\}$$

$$A_0 = z^2[(y_1 - y_1^{-1})(y_2 - y_2^{-1}) - (y_3 - y_3^{-1})^2][(y_1 + y_1^{-1})(y_2 + y_2^{-1}) - (y_3 + y_3^{-1})^2]$$

in which we use the following abbreviations:

$$\left.\begin{aligned} c_{1,2} &= \cosh J_{1,2} \\ s_{1,2} &= \sinh J_{1,2} \\ c_{\pm} &= \cosh (J_2 \pm J_1) \\ s_{\pm} &= \sinh (J_2 \pm J_1) \end{aligned}\right\}$$

Eq.(3.11) can be solved analytically since it is a quartic equation in $\bar{\lambda}$, and the largest root $\bar{\lambda}$ then determined. However the fugacity ratio z appearing in the coefficients still remains unknown. This difficulty can

can be bypassed by utilizing the following relation:

$$X_{AT} \equiv \frac{\langle n \rangle}{N} = \frac{1}{Y} \sum_{n=0}^{N} n e^{-\beta H_e \beta \mu n}$$

$$= \frac{\beta z}{\bar{\lambda}} \frac{\partial \bar{\lambda}}{\partial z} \qquad (3.12)$$

where X_{AT} is the fraction of AT pairs in the DNA. Using Eq.(3.12) z can be eliminated and $\bar{\lambda}$ expressed in terms of X_{AT}, which is a quantity measurable from experiments and which is fixed for a given sample of DNA.

Before proceeding further, it will be helpful to pause briefly for a discussion on the nature of the parameters U and J appearing in the Hamiltonian. In what follows we shall restrict ourselves in the discussion to a three parameter theory in the interest of simplicity: $J_1(T)$ and $J_2(T)$ to characterize the different thermal stabilities of AT and GC pairs, and U to specify the stacking interaction. In general, the nearest neighbour stacking interaction U depends on the nature of the neighbouring base pairs in question, as well as their steric configuration. We shall have at least six different U's before we can take care of all the possibilities. Consequently this would complicate our theory so that only numerical results would be then available. However up to the present time, there has been no experimental evidence for the need to distinguish such fine details, as indicated by the experiments of Brahms et al (1967), where very similar stacking interactions for various bases are suggested. An indication leading to similar conclusions may be seen by measurement of the transition widths of dAT:dAT and dG:dC in 0.001 M Na_2EDTA + 0.005 M NaH_2PO_4 + 0.0025 M Na_2HPO_4 (Montroll et al, 1966). The disagreement for the values of J_1 the two homopolymers is about 10%. We shall assume that all stacking inter-

actions are identical and are also independent of the nature of the base pairs and temperature. This assumption is also used in other theoretical studies (Montroll et al, 1966; Lehman and McTague, 1968).

$$U\binom{AT}{AT} = U\binom{AT}{TA} = U\binom{GC}{AT} = U\binom{GC}{TA} = U\binom{GC}{GC} = U\binom{GC}{CG} = U \quad . \tag{3.13}$$

It is also helpful to define the cooperativity factor σ as

$$\sigma = \exp(-4U) \tag{3.14}$$

The parameters $J_1(T)$ and $J_2(T)$ are temperature dependent. They characterize the different thermal stabilities or free energies of the hydrogen bonds in the AT and GC pairs. They are also related to the relative probabilities of finding a base pair intact or disrupted at a given temperature. It follows therefore that if T_1 and T_2 are the melting temperatures of the AT and GC homopolymers,

$$J_1(T_1) = J_2(T_2) = 0$$

such that $\exp(2J_1(T))_{T=T_1} = \exp(2J_2(T))_{T=T_2} = 1$.

There are at present two major versions for the temperature dependence of J_1 and J_2. Following the arguments of Montroll and Goel (1966), one gets the following:

$$\text{CASE (A)} \qquad J_1(T) = a_1(T_1 - T) \; ; \; J_2(T) = a_2(T_2 - T) \tag{3.15}$$

in which a_1 and a_2 are constants.

The validity of Eq.(3.15) follows from the assumption that the enthalpy and entropy terms in the expression for the free energy are

linear functions of temperature in the neighbourhood of the melting point. In their calculation, Montroll et al (1966) further assume that

$$a_1 = a_2 = a \tag{3.16}$$

However, following Lehman's arguments (Lehman, 1967 and Lehman and McTague, 1968), one arrives at the following:

$$\text{CASE (B)} \quad J_1(T) = b_1(1/T - 1/T_1) \; ; \; J_2(T) = b_2(1/T - 1/T_2) \tag{3.17}$$

where b_1 and b_2 are constants.

If one further assumes that the entropies of melting AT and GC pairs are the same, it follows that

$$b_1/T_1 = b_2/T_2 \tag{3.18}$$

Although the temperature dependence of the J's in these two choices appears to be quite different, within the experimentally meaningful range for T and using proper values for the parameters, the actual deviation of the above two forms is not great (see Fig. 3.2). With reference to later sections, we suggest the following method to resolve the problem created by this ambiguity in the temperature dependence of J_i's. The following relation between $J_1(T)$ or $J_2(T)$ and the experimentally determinable quantities ϕ_1, ϕ_2 and ϕ (defined below in Eqs. (3.26) to (3.28)) can be derived (see Appendix III):

$$J_i(T) = \ell n \left\{ \left(\frac{1 + \phi_i}{1 - \phi_i} \right)^{\frac{1}{2}} \cdot \frac{\sigma^{\frac{1}{2}}\phi + \sqrt{1 - (1 - \sigma)\phi^2}}{1 + \phi} \right\} \; ; \; i = 1,2 \tag{3.19}$$

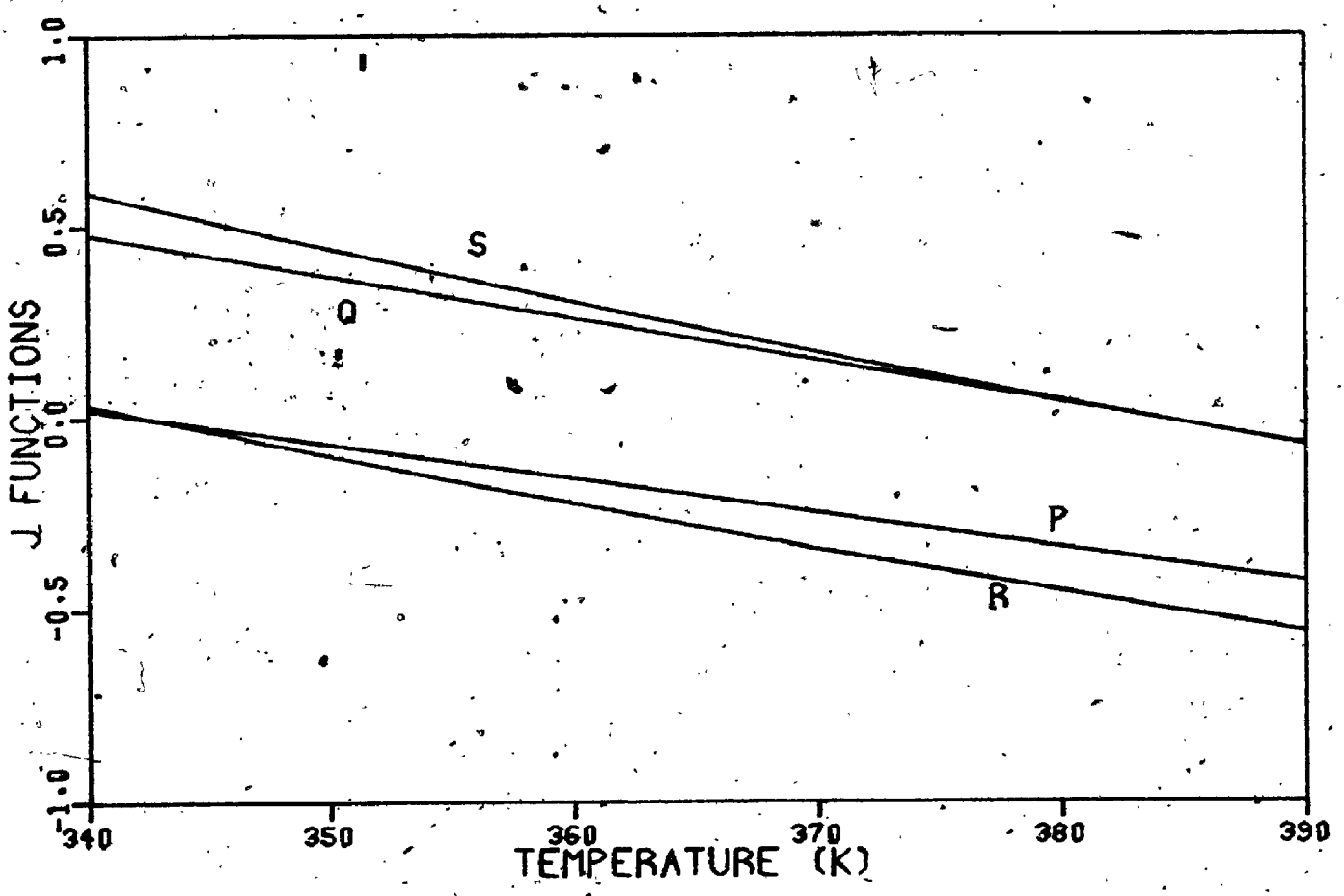


FIGURE 3.2 COMPARISION BETWEEN THE TWO FUNCTIONAL FORMS OF J: CASE(A). CURVES P AND Q. CASE(B). CURVES R AND S.

The right side of Eq. (3.19), and therefore $J_i(T)$, can be determined as a function of temperature experimentally (since ϕ_1, ϕ_2 and ϕ are experimentally determinable), using the proper value for the cooperativity factor σ.

Returning to the discussion on the theory itself, it follows from Eq. (3.13) that the last two terms in Eq. (3.11) vanish and consequently the largest eigenvalue of Q, $\bar{\lambda}$, (hereafter written as λ) becomes

$$\lambda = y(c_1 z + c_2 + R) \tag{3.20}$$

where

$$R = \{ (s_1 z + s_2)^2 + \sigma(z^2 + 2c_- z + 1) \}^{\frac{1}{2}} \tag{3.21}$$

and

$$X_{AT} = z\{ c_1\lambda - y(1-\alpha)(z + c_-) \}/(R\lambda) \tag{3.22}$$

The quantities $\theta(T)$ (the fraction of disrupted base pairs), $\theta_1(T)$ (the fraction of disrupted AT pairs) and $\theta_2(T)$ (the fraction of disrupted GC pairs) can be expressed as follows according to their definitions:

$$\theta = < \tfrac{1}{2} \sum_{j=1}^{N} (1 - \tau_j) > = \tfrac{1}{2} \{ 1 - (s_1 z + s_2)/R \} \tag{3.23}$$

$$\theta_1 = \tfrac{1}{2} \{ 1 - \frac{yz}{X_{AT}\lambda} (s_1 + \frac{s_1(c_1 z + c_2) + (1 - \sigma)s_-}{R}) \} \tag{3.24}$$

$$\theta_2 = \tfrac{1}{2} \{ 1 - \frac{y}{(1 - X_{AT})\lambda} (s_2 z + \frac{s_2(c_1 z + c_2) + (1 - \sigma)s_- z}{R}) \} \tag{3.25}$$

Now define the quantities

$$\phi = 1 - 2\theta = (s_1 z + s_2)/R \tag{3.26}$$

$$\phi_1 = 1 - 2\theta_1 = \{s_1\lambda + y(1-\sigma)s_-\}/\{c_1\lambda - y(1-\sigma)(z + c_-)\} \quad (3.27)$$

$$\phi_2 = 1 - 2\theta_2 = \{s_2\lambda - y(1-\sigma)s_-\}/\{c_2\lambda - y(1-\sigma)(zc_- + 1)\} \quad (3.28)$$

It is clear from their definitions above that

$$\theta = X_{AT}\theta_1 + (1 - X_{AT})\theta_2 \quad (3.29)$$

and

$$\phi = X_{AT}\phi_1 + (1 - X_{AT})\phi_2 \quad (3.30)$$

Equations (3.20)-(3.22), (3.26)-(3.28) form a set of six independent coupled algebraic equations in the variables λ, z, R, ϕ_1, ϕ_2 and ϕ. It is possible to solve for all of the six quantities as functions of the independent variable T. For example we may derive the following quartic equation for ϕ:

$$B_4\phi^4 + B_3\phi^3 + B_2\phi^2 + B_1\phi + B_o = 0 \quad (3.31)$$

where

$$B_o = -\Delta^2$$

$$B_1 = -2\Delta s_1 s_2$$

$$B_2 = \Delta^2 - s_1^2 s_2^2 + \sigma c_+^2 - 2\sigma s_+$$

$$B_3 = 2(\Delta s_1 s_2 + \sigma c_1 c_2 s_+ - \sigma\Delta c_+)$$

$$B_4 = (s_1^2 + \sigma)(s_2^2 + \sigma)$$

and

$$\Delta = c_1 s_2 - X_{AT} s_- \quad (3.32)$$

Note that all general polynomials of degree four or less can be solved analytically, Eq. (3.31) being no exception. The solution for Eq. (3.31) is nevertheless too complicated to be displayed. However a number of

interesting results can be derived even without actually carrying out fully the solving of the equation. They shall be presented in, Chapter IV.

CHAPTER IV

APPLICATION TO NATURAL DNA

In order to give a convenient description of the theory, it is important to determine the independent and dependent variables. For an independent variable, the temperature T is an obvious choice. We also choose the fractional AT content X_{AT} (hereafter written as X) as an adjustable parameter for the reason that it characterizes a given DNA molecular species to a certain extent and varies from one species to another. An extra advantage is that X can be easily determined from experiments. All other quantities are then expressible in terms of T and X.

The general formulation of the theory has been presented previously in Chapter III. Quartic equations with T and X dependent coefficients in each of the earlier defined variables ϕ, z, λ, h, γ and R have been obtained (see Appendix I). Among these equations, Eq. (3.31) is of particular importance since it most conveniently describes the behaviour of the melting curve through the relation

$$\theta(T) = \tfrac{1}{2}\,(1 - \phi(T)) \qquad (4.1)$$

so that a solution of Eq. (3.31) gives $\theta(T)$ directly. Fig. 4.1 gives a comparison between the experiment and our present theory (De Ley and Schell, 1963). Eq. (3.31) may also be used to predict the behaviour of the melting curve at the melting temperature T_m, when $\phi = 0$

$$\Delta(T_m) = c_1 s_2 - X s_- \Big|_{T=T_m} = 0$$

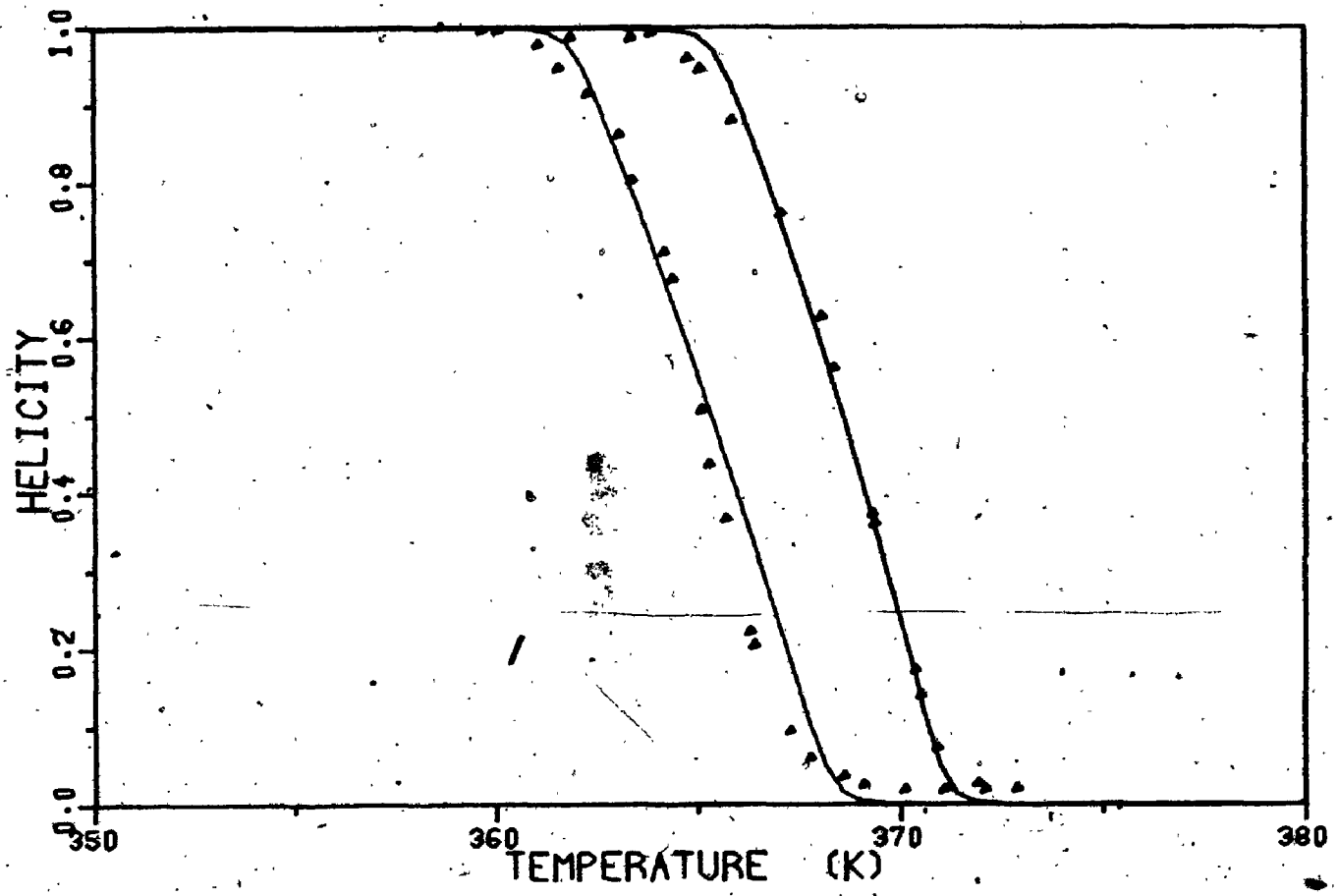


FIGURE 4.1 MELTING CURVES OF (LEFT) A. ACETI. "PARADOXUS" P2 AND (RIGHT) A. ACETI. "LIQUEFACIENS" 28.
▲ DATA OF DE LEY AND SCHELL
SOLID LINE IS THEORETICAL CURVE USING $a_1 = 0.007$ K. $a_2 = 0.009$ K.
$T_1 = 340$ K. $T_2 = 381$ K.
$\alpha = 5 \times 10^{-5}$

which can be written in the alternative form

$$\frac{\tanh J_1(T_m)}{\tanh J_2(T_m)} = -\frac{1-X}{X} \qquad (4.2)$$

Eq. (4.2) in general represents a transcendental relation between X and T_m. However, substituting correct values for the parameters in J_1 and J_2, this relation turns out to be almost linear (see Fig. 4.2). In fact, applying the appropriate values for J_1 and J_2 determined from independent experiments involving enthalpies to fit the data reported by Marmur and Doty (1962) and by De Ley (1970), the result is found to be very favourable (Leung and Tong). Eq. (4.2) yields as a limiting case the linear empirical formula first proposed by Marmur and Doty (1962):

$$T_m = T_2 - X(T_2 - T_1) \qquad (4.3)$$

It is clear from Eq. (4.2) that the melting temperature of heterogenous DNA depends on X, J_1 and J_2 and, rather unexpectedly, not at all on the cooperativity factor σ. This independence of T_m on σ renders Eq. (4.2) an excellent means of estimating the coefficients in the expressions for J_1 and J_2. Using case (A), it was found that the set

$$a_1 = 0.009K^{-1}, \quad a_2 = 0.011K^{-1}, \quad T_1 = 342.5K, \quad T_2 = 383.5K$$

gives a good fit in Fig. 4.2. Montroll et al used the numbers

$$a_1 = a_2 = 0.01K^{-1}, \quad T_1 = 340K, \quad T_2 = 381K.$$

Using case (B) for J_1 and J_2 and making the assumptions that the

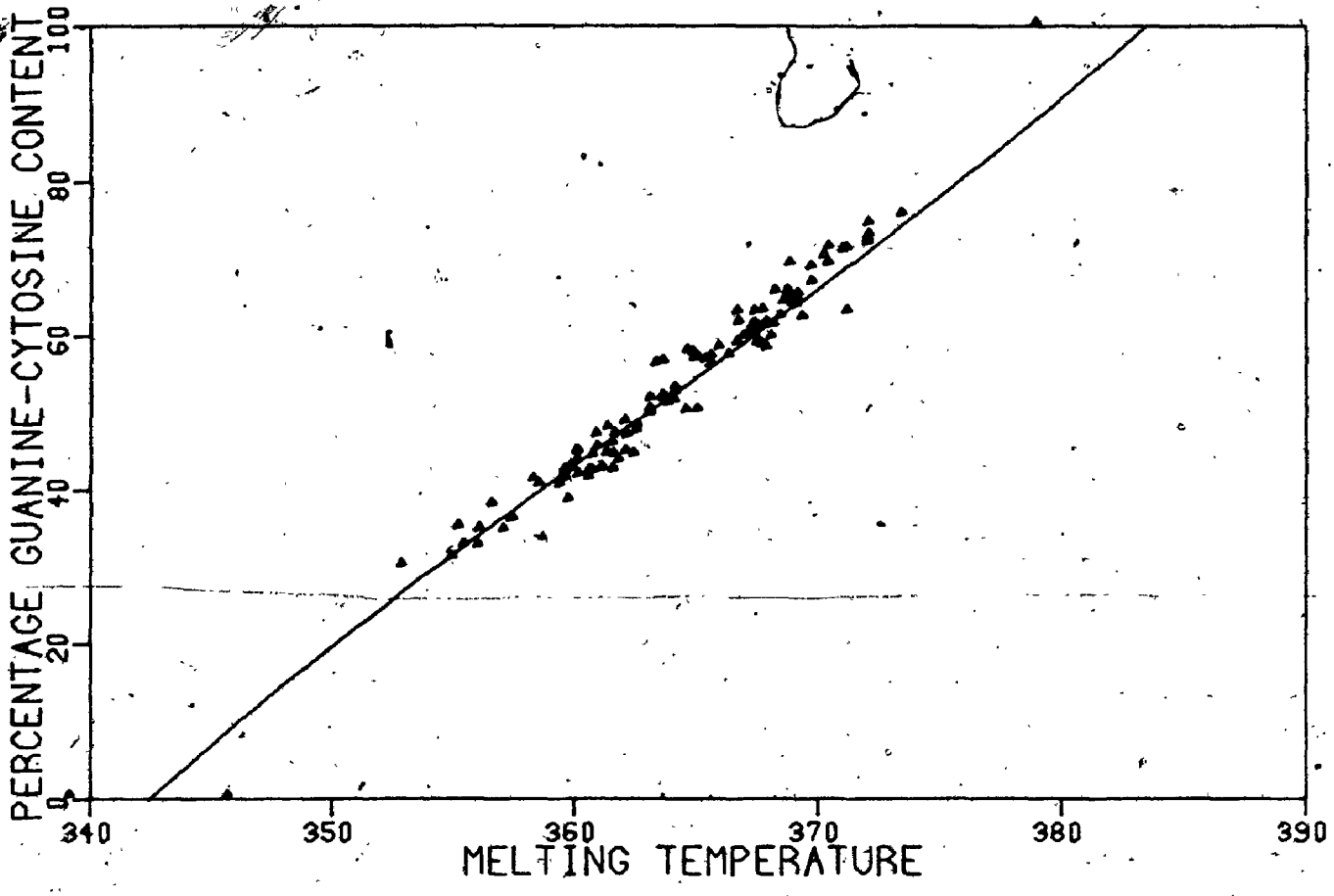


FIGURE 4.2 VARIATION OF MELTING TEMPERATURE WITH PERCENTAGE GC CONTENT
△ DATA OF DE LEY.
SOLID LINE IS THEORETICAL CURVE USING b_1 = 1900 K, b_2 = 2127 K.
T_1 = 342.5 K; T_2 = 383.5 K.

TABLE 4.1

a_1 /K	a_2 /K	T_1 K	T_2 K	S.D.x100
0.0100	0.0100	342.5	383.5	2.23
0.0055	0.0075	337.0	380.5	2.14
0.0055	0.0075	338.5	381.0	2.13
0.0057	0.0080	338.1	380.8	2.13
0.0060	0.0080	339.0	381.0	2.14
0.0070	0.0090	340.0	381.0	2.15
0.0075	0.0100	339.0	381.0	2.14
0.0080	0.0115	338.0	380.5	2.14
0.0090	0.0120	339.0	381.0	2.14
0.0095	0.0115	340.5	381.5	2.15

TABLE 4.2

b_1 K	b_2 K	T_1 K	T_2 K	S.D.x100
900	1400	338.6	380.7	2.14
1000	1600	338.2	380.5	2.14
1100	1700	338.7	380.7	2.14
1200	1900	338.6	380.4	2.14
1573	1762	342.5	383.0	2.27
1910	2140	342.5	383.0	2.23
1900	2127.4	342.5	383.5	2.30

entropy of formation for AT and GC pairs are the same and that the difference $T_2 - T_1$ is 41K in SSC, the set $b_1 = 1573.6K$, $b_2 = 1762.0K$, $T_1 = 342.53K$ and $T_2 = 383.53K$ yielded the best results. Compare these with Lehman's set: $b_1 = 1900K$, $b_2 = 2127.4K$, $T_1 = 342.5K$ and $T_2 = 383.5K$. Tables 4.1 and 4.2 give the values of T_1, T_2, a_1, a_2 and b_1, b_2 that give a better fit to De Ley's data. Once these parameters are determined it becomes a matter of simple calculation to find the fractional AT content X of an unknown sample of DNA using Eq. (4.2) if the melting temperature T_m is available from experiments. These experiments can be carried out by a variety of techniques. The determination of X is a subject which has been much discussed in biology. This quantity X is one of the very few in biology that can be subjected to quantitative treatment, and therefore it is particularly valuable in bacterial taxonomy, where it has shed new light on problems such as bacterial classification and genetic relatedness among species. For a detailed discussion on the bearing of X upon bacterial taxonomy we refer the reader to the article by Mendel (1969).

We shall present another method for the determination of X. Using the present scheme of calculation, we can derive an expression relating X and any conveniently fixed temperature T_o. For example, let us choose $T_o = 363K$ for SSC conditions. This corresponds to the melting temperature of a DNA with 50% AT content. Then we find that

$$X = \tfrac{1}{2}\left\{1 - \frac{\phi_o}{c_o s_o}\left(s_o^2 + \frac{-\sigma\phi_o^2 + \sqrt{\sigma(1-\phi_o^2+\sigma\phi_o^2)}}{1-\phi_o^2}\right)\right\} \qquad (4.4)$$

where

$$s_o = \sinh J_2(T_o),\ c_o = \cosh J_2(T_o),\ \phi_o = 1 - 2\theta(T_o)\ .$$

$\theta(T_o)$ is the value of the fraction of disrupted base pairs at the temperature T_o (Leung and Tong). With a prior knowledge of s_o, c_o and σ, we only need to take one accurate measurement of the normalised melting curve to find $\theta(T_o)$ at the fixed preassigned temperature (363 K in this case) and X can then be quickly evaluated from Eq.(4.4). The plot of X versus the fraction of intact base pairs (helicity) at $T = T_o$ is presented in Fig. 4.3. This curve seems rather insensitive to the value of the cooperativity factor σ. This curve is most useful if the given DNA has X values between 0.35 and 0.65, because the corresponding $\theta(T_o)$ values almost range from 0 to 1, so that roughly a 3% change in X would result in a 10% change in $\theta(T_o)$.

In the present method, the transition width W can be derived through evaluating the temperature derivative of θ at the melting temperature. A careful examination of Eq.(3.31) reveals that there are two roots which vanish simultaneously at T_m. Therefore extra care must be taken to keep track of the physical root and its derivative in the neighbourhood of T_m. For the two functional forms of J we find

$$\text{(A)} \qquad W = \frac{2(\sigma^{\frac{1}{2}}c_+ - s_1 s_2)}{a_1(Xc_- + s_1 s_2) + a_2(c_1 c_2 - Xc_-)} \qquad \text{at } T = T_m \qquad (4.5A)$$

$$\text{(B)} \qquad W = \frac{2T_m^2(\sigma^{\frac{1}{2}}c_+ - s_1 s_2)}{b_1(Xc_- + s_1 s_2) + b_2(c_1 c_2 - Xc_-)} \qquad \text{at } T = T_m \qquad (4.5B)$$

In both forms the width varies as $\sigma^{\frac{1}{2}}$ (see Fig. 4.4). This observation differs from that concluded by other authors (Vedenov et al, 1972; Lehman 1967; Shagulii et al, 1969), who derive a logarithmic dependence of W on σ. Nonetheless, the difference may very well be attributed to

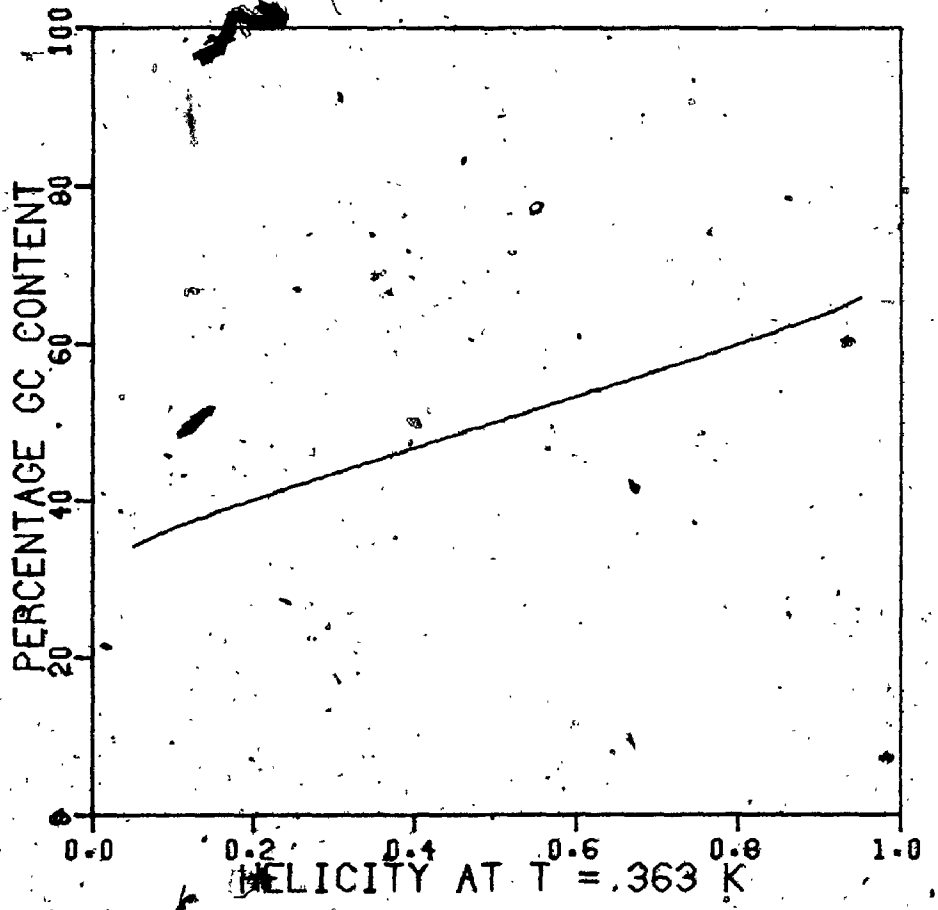


FIGURE 4.3 DETERMINATION OF PERCENTAGE GC CONTENT BY MEASURING HELICITY AT 363 K

b_1 = 1900 K, b_2 = 2127 K,

T_1 = 342·5 K T_2 = 383·5 K,

$\sigma = 5 \times 10^{-5}$

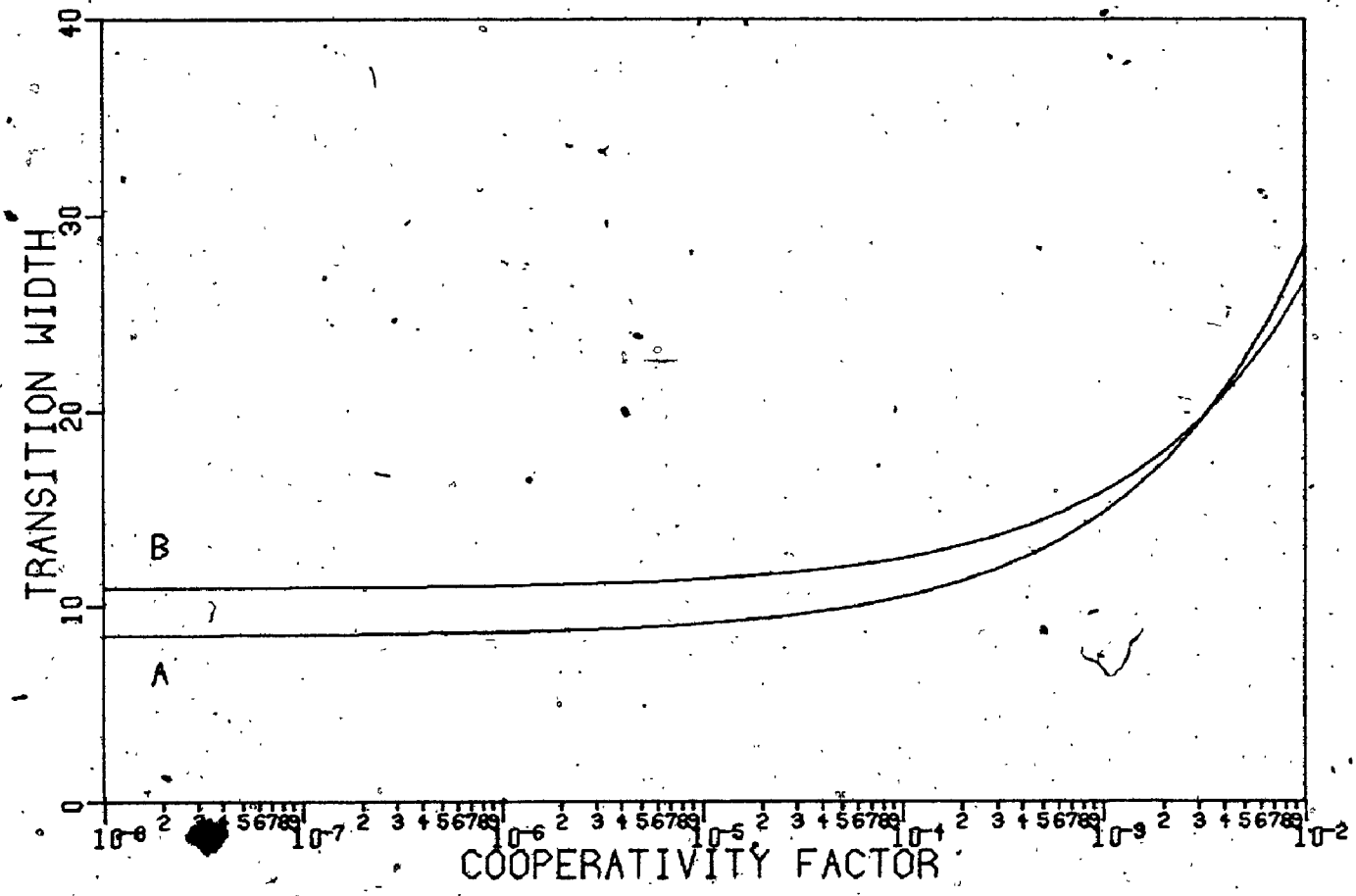


FIGURE 4.4 VARIATION OF TRANSITION WIDTH WITH COOPERATIVITY FACTOR: FOLLOWING (A) EQN. 4.5A AND (B) EQN. 4.5B.

the adopted definitions for W. The dependence on X is not so simple, because now the hyperbolic functions are all evaluated at T_m, which in turn is related to X through Eq.(4.2). As is evident from Fig. 4.5, the overall variation with X is very much parabolic. The peak value centres around X = 0.5 and drops sharply for the case of homopolynucleotides. In SSC, the value of W ranges from a fraction of a degree for homopolynucleotides to about 10° for DNA's having 50% AT content.

The determination of W from experiment is simple but also rather unreliable. This is mainly due to the inherent difficulty in measuring the slope of the melting curve at the transition temperature. A given error in assessing the transition temperature will inevitably result in a larger error for the transition width W.

Another interesting feature associated with the denaturation of heterogeneous DNA is the partial melting curves, denoted hereafter by $\theta_1(T)$ and $\theta_2(T)$, (for AT and GC base pairs respectively). These quantities represent the fraction of a particular base pair type (AT or GC) that has been denatured at a certain temperature; whereas $\theta(T)$ described earlier represents the overall fraction of denatured base pairs, irrespective of whether the pair is AT or GC. The study on the partial melting curves is an area that has not been covered sufficiently in the literature. Felsenfeld and Sandeen (1962) were among the first to point out that because AT and GC base pairs make different contributions to the absorption spectrum, a careful analysis of the spectrum would therefore give information about these partial melting curves. Fig. 4.6 illustrates the shapes of these curves and their relationship with the overall melting curve $\theta(T)$. It is important to note that the temperature

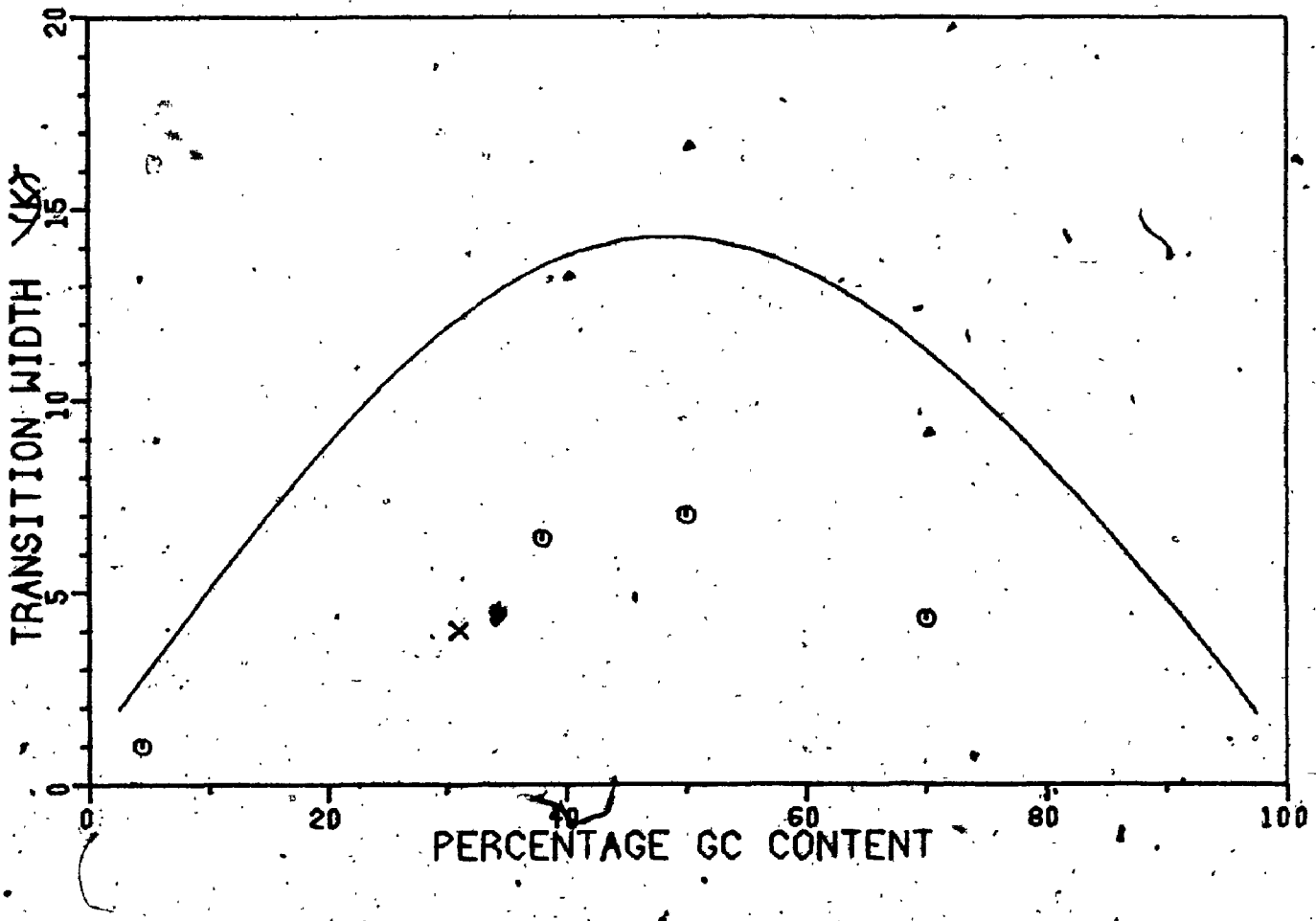


FIGURE 4.5 VARIATION OF TRANSITION WIDTH WITH PERCENTAGE GUANINE-CYTOSINE CONTENT

DATA OF DAVIDSON ET AL:

● USING 0.1M $NaClO_4$

△ USING $3X10^{-4}$M $NaClO_4$

× DATA OF CROTHERS ET AL. USING 0.018M NaCl.

SOLID LINE IS THEORETICAL CURVE USING $b_1 = 1900$ K. $b_2 = 2140$ K. $T_1 = 342.5$ K. $T_2 = 383.5$ K. $\sigma = 5 \times 10^{-5}$

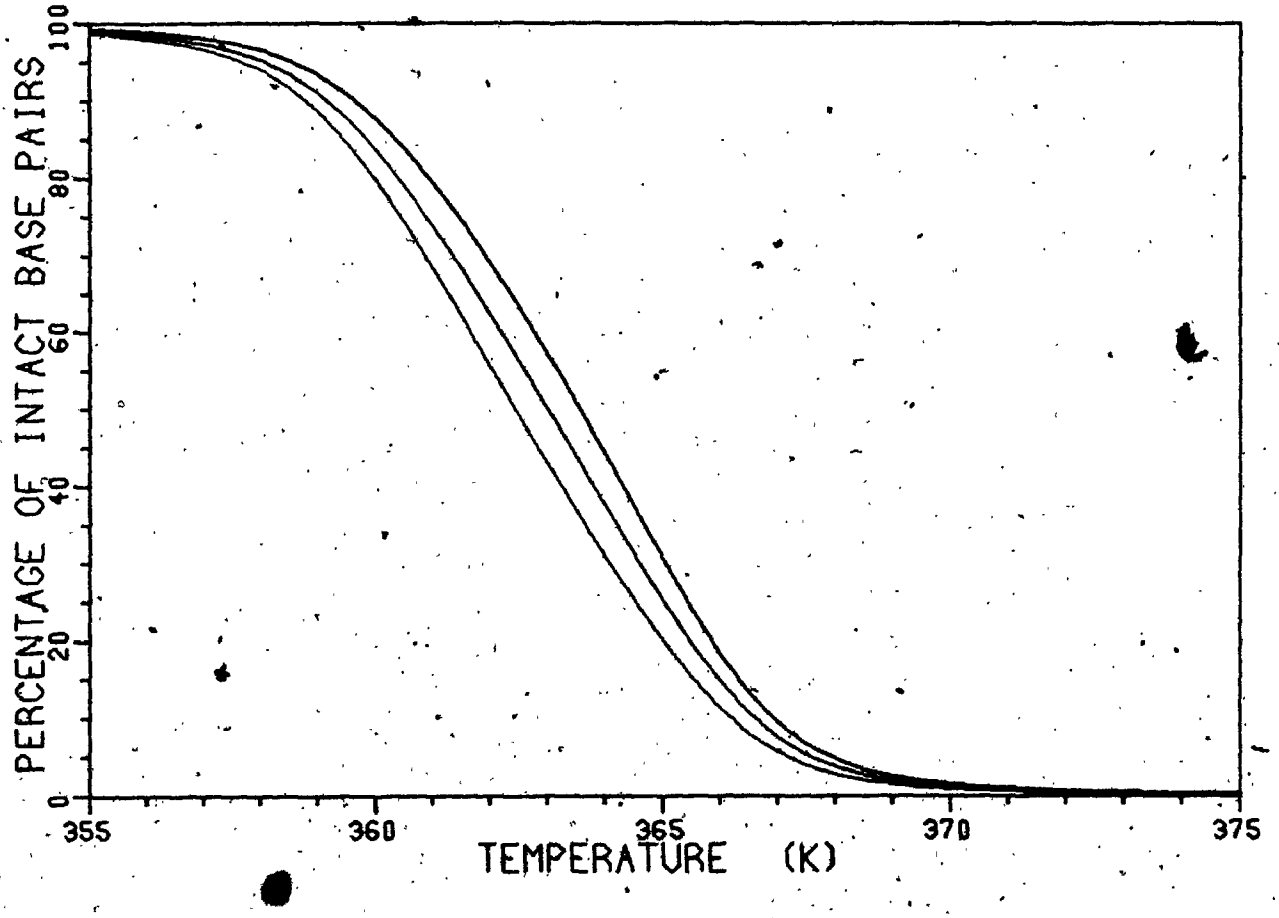


FIGURE 4.6 AT PARTIAL (LEFT), GC PARTIAL (RIGHT) AND OVERALL MELTING CURVES OF E. COLI AT SSC, USING $a_1 = 0.006\ K^{-1}$, $a_2 = 0.008\ K^{-1}$, $T_1 = 339\ K$, $T_2 = 381\ K$, $\sigma = 5 \times 10^{-5}$

separation between these curves is about an order of magnitude narrower than that between the individual melting curves of the AT and GC homopolymers. This immediately suggests that these partial curves are highly sensitive to sequence effects and possibly also cooperativity. In terms of the fugacity ratio z and X, the variables ϕ, ϕ_1 and ϕ_2 can be written:

$$\phi_1 = \{(1 - X)z - Xc_-\}/Xs_- \tag{4.6}$$

$$\phi_2 = \{(1 - X)c_- - X/z\}/(1 - X)s_- \tag{4.7}$$

and $$\phi = \{(1 - X)z + (1 - 2X)c_- - X/z\}/s_- \tag{4.8}$$

If ϕ as a function of temperature is known, then the temperature dependence of the partial curves can be determined through the following equations:

$$\frac{1 + \phi_\ell}{1 - \phi_\ell} = \exp(-2J_\ell(T)) \left(\frac{\sigma^{\frac{1}{2}}\phi + \sqrt{1 - \phi^2 + \sigma\phi^2}}{1 + \phi} \right)^2 \tag{4.9}$$

$\ell = 1, 2$. (See Appendix III for details.)

Because of the sensitive dependence on sequence properties, $\theta_1(T)$ and $\theta_2(T)$ can be used to investigate intramolecular heterogeneity and base clustering in DNA. This was what Falkow and Cowie (1969), and Hirschman and coworkers (1967) did to study bacterial DNA. There, instead of studying these partial curves separately, the direct relation between θ_1 and θ_2 was examined. Fig. 4.7 gives a comparison of our calculation with the data of Hirschman et.al. It is easily seen from such plots that the slope at any point multiplied by the overall AT to GC content ratio would give the ratio of AT to GC pairs that is just denaturing at that

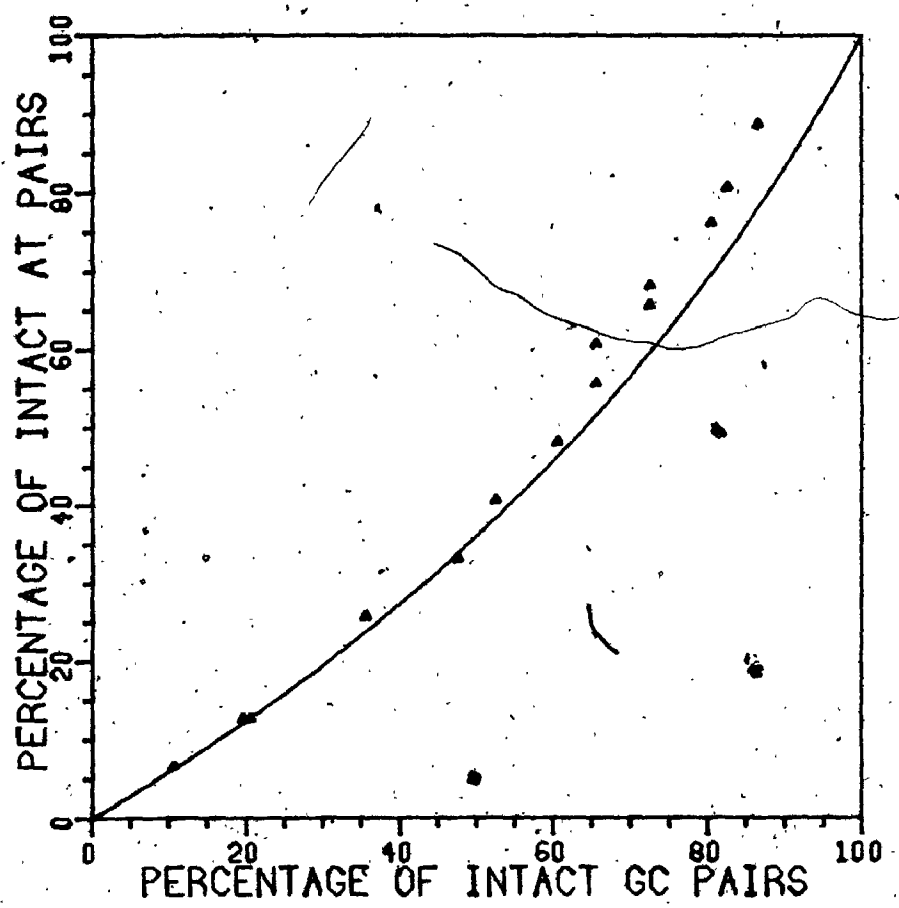


FIGURE 4.7 COMPARISON OF THEORY AND EXPERIMENT ON INTACT AT AND GC BASE PAIRS IN THE TRANSITION PROCESS

△ DATA OF HIRSCHMAN ET AL

— THEORY $a_1 = 0.006\ K^{-1}$, $a_2 = 0.008\ K^{-1}$

$T_1 = 339\ K$, $T_2 = 381\ K$, $\sigma = .5 \times 10^{-5}$

particular temperature corresponding to that point. Thus

$$\frac{X \quad \Delta\theta_1}{1 - X\ \Delta\theta_2} = \frac{\text{Increase of disrupted AT pairs in temp. interval } \Delta T}{\text{Increase of disrupted GC pairs in temp. interval } \Delta T}$$

According to this interpretation and the reasoning that the observed slope decreases monotonically, we infer that the process of denaturation does not proceed uniformly throughout the molecule, but rather, regions of higher AT content tend to denature first at a lower temperature. The agreement between the theoretical curve from our calculation and experimental data is remarkably close.

From Eq.(4.8) the following formula may be derived:

$$\frac{\theta_1(T)}{1 - \theta_1(T)} = k(T) \frac{\theta_2(T)}{1 - \theta_2(T)} \qquad (4.10)$$

in which $k(T) = \exp 2\{J_2(T) - J_1(T)\}$

Through the function $k(T)$, we are now in a position to express the ratio of disrupted to intact base pairs of the AT type to that of the GC type. This function $k(T)$ is of special interest since it is a measure of the ratio of the statistical weights of the two base pair types as used in Lehman's calculation. In their paper, Fink and Crothers (1968) gave a comparison between the various theoretical models. In that paper the statistical weights of AT and GC pairs denoted by s_A and s_B in Lehman's calculation are related to the present J_1 and J_2 in the following way:

$$s_A = \exp(-2J_1) \quad , \quad s_B = \exp(-2J_2) \qquad (4.11)$$

so that

$$k(T) = \frac{s_A}{s_B} = \frac{\theta_1(T)/(1 - \theta_1(T))}{\theta_2(T)/(1 - \theta_2(T))} \qquad (4.12)$$

Eq.(4.12) provides the linkage between observable quantities on the right hand side and the theoretical parameters s_A and s_B or J_1 and J_2. One way to utilize the equation is to estimate the parameters $a_{1,2}$ or $b_{1,2}$. This equation may also be used to establish the temperature dependence of J_1 and J_2. To illustrate this, Fig. 4.8 gives a plot of k(T) as given by Eq. (4.12), using the data of Felsenfeld et al (1962). Fig. 4.9 gives the plot of the logarithm of k(T) and this is precisely $2(J_2 - J_1)$. If J_1 and J_2 depend on temperature in a similar manner, this dependence (whether it is T, 1/T, or some other simple functional form) could be observed by studying Fig. 4.9.

Of special interest to us is another set of observables known as the mean helical length (denoted by h(T)) and mean length of a helix-plus-coil region (denoted by $\gamma(T)$). Both of these quantities are dimensionless, being measured in terms of number of base pairs: h(T) being the average length of a helical region during the transition and $\gamma(T)$ being the average distance between consecutive initiation points of helical regions. In other words, $\gamma(T)$ is the average number of base pairs in a unit cell of the superlattice of helix and coil regions. As we shall see later, these quantities, in contrast to the transition width, are very sensitive to the cooperativity factor σ, particularly when evaluated at the transition temperature, where $h(T_m)$ and $\gamma(T_m)$ depend on σ as $\sigma^{-\frac{1}{2}}$. This property renders them very valuable in the

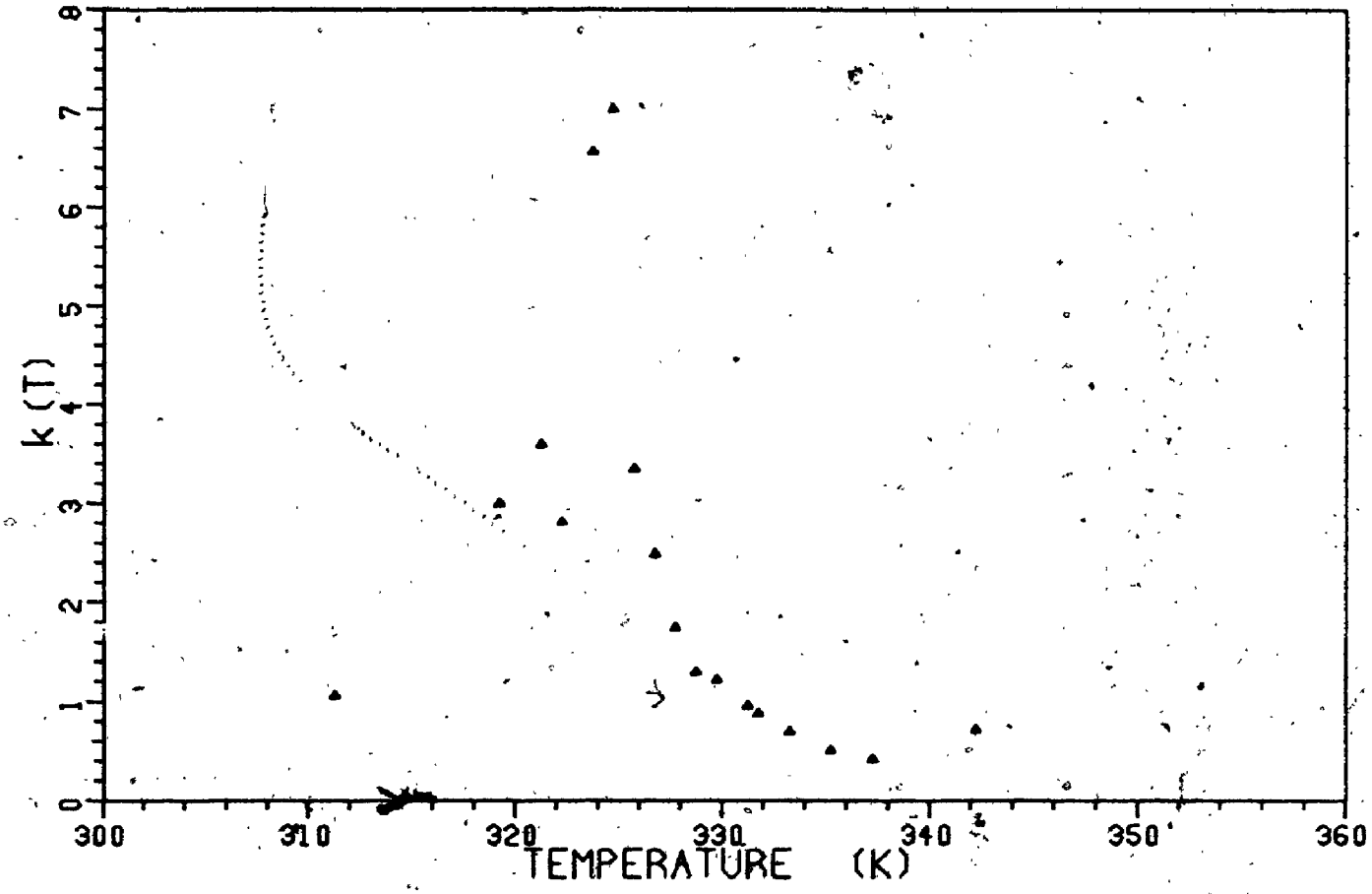


FIGURE 4-8 PLOT OF k(T) AGAINST TEMPERATURE
△ DATA OF FELSENFELD ET AL

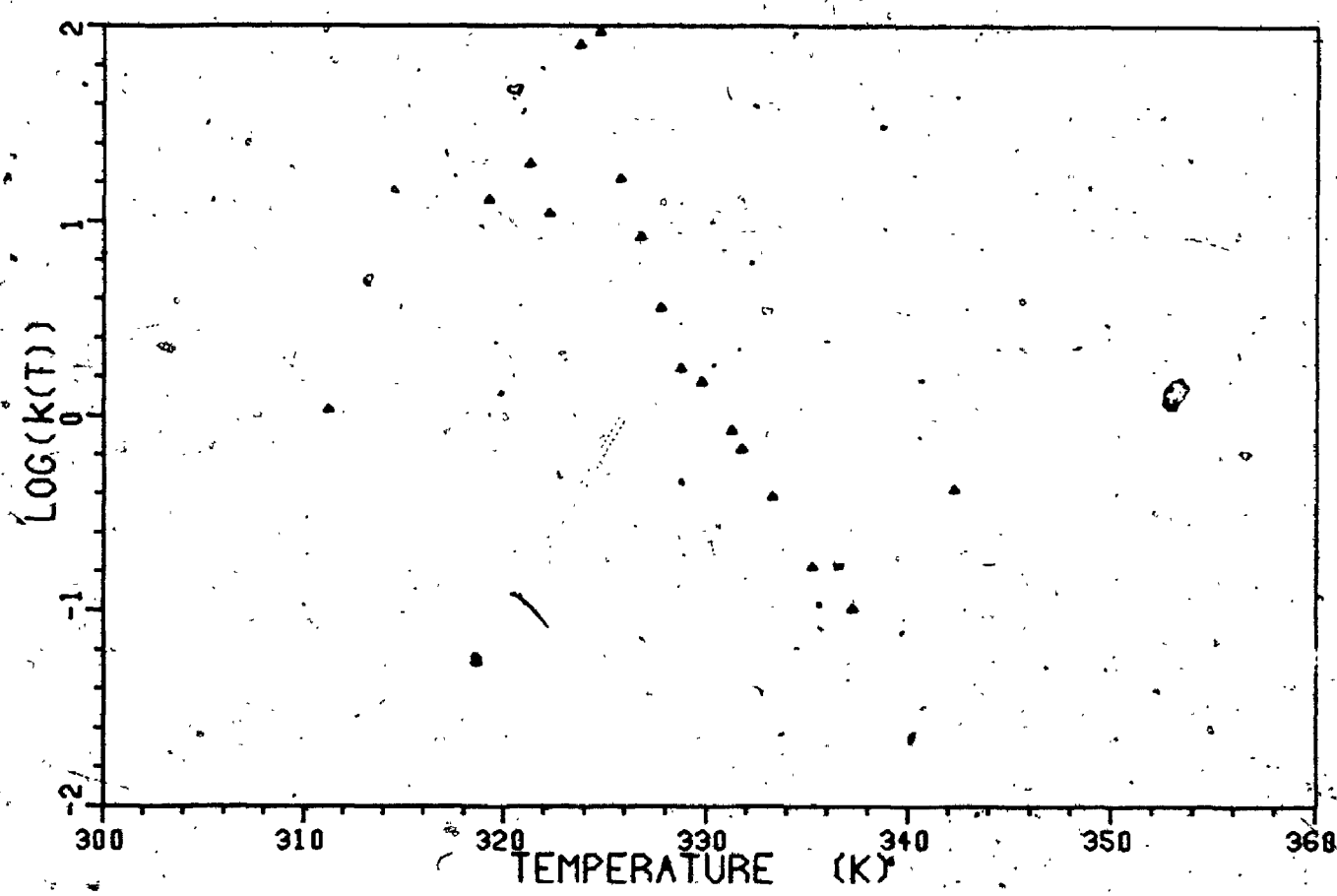


FIGURE 4.9 LOG(K(T))
△ , DATA OF FELSENFELD ET AL
T_m = 330 K

determination of σ. Currently, a rough estimate for σ is commonly quoted as 10^{-6}. Unfortunately experimental information related to σ is extremely scarce. A viscosimetric investigation of T2 phage DNA (X = 0.65) was reported by Shagulii et al (1969), where owing to the presence of other complications in the later stages of denaturation, sufficiently reliable data could only be obtained for θ values between 0.1 and 0.2.

The dependence of h and γ on the degree of transition φ can be derived directly from the present theory (see Appendix IV). We find

$$h = 1 + \frac{\phi + \sigma^{-\frac{1}{2}}\sqrt{1 - \phi^2 + \sigma\phi^2}}{1 - \phi} \qquad (4.13)$$

and

$$\gamma = \frac{2(1 + \sigma^{-\frac{1}{2}}\sqrt{1 - \phi^2 + \sigma\phi^2})}{1 - \phi^2} \qquad (4.14)$$

Of special interest are the values of h and γ at the transition temperature, where $\phi = 0$; we find directly:

$$h(T_m) = 1 + \sigma^{-\frac{1}{2}} \qquad (4.15)$$

$$\gamma(T_m) = 2(1 + \sigma^{-\frac{1}{2}}) = 2h \qquad (4.16)$$

It can be proved that Eq. (4.16) represents the mathematical minimum value of γ. One can therefore utilize Eq. (4.15) or (4.16) to find the cooperativity factor σ. As $\sigma \sim 10^{-6}$ in SSC, $\sigma^{-\frac{1}{2}} \sim 10^3$. It follows that even an order-of-magnitude estimation of h might turn out to be sufficient for a good estimation of σ. The advantage of this method is further amplified by the fact that the log h versus θ plot is very slowly varying in the neighbourhood of the melting temperature (see Fig. 4.10, so that the order-of-magnitude estimation will not be far off even if one has

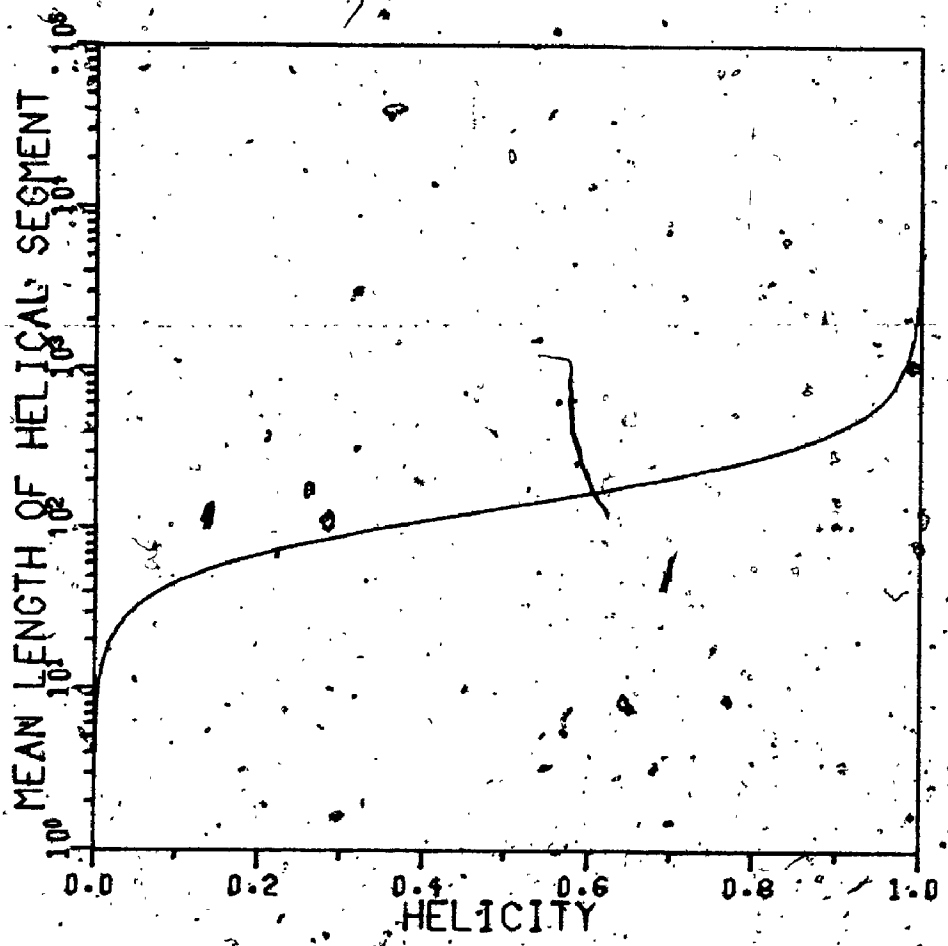


FIGURE 4.10 VARIATION OF MEAN LENGTH OF HELICAL SEGMENTS WITH HELICITY

$b_1 = 1900$ K, $b_2 = 2127$ K.

$T_1 = 342.5$ K, $T_2 = 383.5$ K

$\sigma = 5 \times 10^{-5}$

difficulty in locating the melting temperature.

CHAPTER V

DISCUSSION

A theoretical account on the helix-coil transition of DNA molecules has been presented in preceding chapters. The practicability of such a presentation rests essentially on the validity of the following approximation. The base sequence of a natural DNA molecule, unknown as it still is, is approximated by a sequence that is statistically most probable. The approximant is taken to be the ensemble average of all possible sequences having a GC content consistent with that of the natural sequence. This presently seems to be a reasonable manoeuvre in view of the fact that we are still not in a position to completely specify the entire sequence of a DNA molecule, which typically consists of 10^5 base pairs. All physical properties calculated are based on this averaged sequence. There is no a priori justification that such a scheme would produce the desired properties of the natural DNA, whose sequence is obviously specific (though unknown) in the sense that it embodies the genetic code.

An alternative approximant to the natural sequence, as adopted in some calculations (Fink and Crothers, 1968) is the random sequence. A computer routine is employed to generate a random sequence with probabilities of occurrence of AT and GC pairs proportional to the relative base content of the DNA in question. The inferences drawn therefrom can then be satisfactorily applied to natural DNA's if the natural DNA has a sequence which is not too much different from that particular randomly generated sequence, or if the properties in question

are not particularly sensitive to or entirely independent of the details of base sequencing.

An important conceptual distinction between these two methods of approximation, namely, the most probable sequence and the random sequence should be explained. The random sequence is relatively simple to comprehend. Once it has been generated by the computer, it is explicitly fixed for the remainder of the calculation. On the other hand, the statistically most probable sequence may change, in principle, from one temperature to another. The result is that at any one temperature, only the averaged or statistically most likely properties are evaluated. These may therefore deviate from the corresponding properties of a fixed sequence.

The approach demonstrated by Lehman (1967), Poland et al (1969), Reiss et al (1966), Cohen et al (1970) and the Russian group (Lazurkin et al, 1970; Vedenov et al, 1969) assigns a statistical weight for each site to the partition function. This statistical weight is dependent on the type of base pair that is occupying that particular site. This kind of approach can be solved only formally and the final results must be obtained from numerical calculation. Hence the exact relation between the physical properties and the energy parameters $J_{1,2}$ and U is lost. Although our model may have the conceptual difficulty of having different statistically most probable sequences at different temperatures it does have the advantage of allowing us to present explicit analytical expressions for various measurable quantities. For example, we can now see in what way melting temperatures, correlation lengths, transition widths, etc. are related to the hydrogen bond energies and the stacking

interaction energies of neighbouring base pairs. We also know which properties are most sensitive to which energy parameter, $J_{1,2}$ or U. The transition temperature, for example, is independent of the stacking interaction and therefore is independent of the actual base sequencing. The same result of T_m should be obtained here as in other methods based on specific sequences. This explicit independence of T_m on stacking interaction (or the cooperativity to which it is related) agrees well with the experimental finding that transition temperatures are very stable and sequence independent. Expressions for the transition width, mean length of helical regions and helix-plus-coil regions and their dependence on temperature, GC content, helicity or the energy parameters have also been explicitly obtained. It is observed that our calculation generally predicts transition widths that are slightly broader than those calculated by Lehman (Lehman 1967; Lehman and McTague, 1968). These explicit expressions have an obvious advantage over the numerical results in that while numerical output tends to obscure the important underlying factors so that full advantage of one's physical insight cannot be utilized, explicit expressions are easier to work with and usually enhance the possibility for deeper interpretation and further manipulation.

It remains to point out that despite the fact that we have used a grand partition function, the present calculation is quite different from the usual canonical ensemble calculations (Huang, 1963; Wannier, 1966, Zimm, 1960; Lifson and Zimm, 1963). The difference dwells in the total number of base pairs N. In the present work this parameter is held fixed (and assumed to be large) while the relative amounts of the two species of base pairs may vary. However in conventional calculations the

parameter N corresponding to the number of sites in a statistical ensemble usually changes.

Our calculation gives a general description of the transition process and related phenomena from a unified viewpoint. A consistent set of parameters has been used throughout the exposition. They are no longer treated as separate entities related only phenomenologically. In a sense the theory itself is a step towards a more fundamental understanding of the transition process, as it is now explanatory in terms of molecular processes that take place among the base pairs, rather than in terms of empirical observations. Previously related only phenomenologically, they now become integral parts of a common framework. This will be an asset for future work and perhaps also for studies on processes such as DNA unwinding.

It is encouraging to see that the values of the parameters, determined from independent sources, do not vary throughout our calculation for various situations. For example, the values of b_1 and b_2 obtained from enthalpy measurements reduce the transcendental Eq.(4.2) almost to a linear relation, in close agreement with the apparently unrelated set of data on the dependence of T_m on X. This is a good indication that the present theory is on the right track.

The values for the various parameters used earlier must be understood to be tentative. We are sure that they can be improved as more and more experimental information is made available. We are badly in need of a more systematic and detailed measurement of the overall and partial melting curves and the transition widths of different species of DNA under identical

solvent conditions. Also lacking is experimental information on the mean length of a helical region during transition, especially in the neighbourhood of the melting temperature. The magnitude of the stacking energy in vivo is still uncertain. We know that the value of the mean length of a helical segment depends strongly on the cooperativity factor σ (and therefore the stacking energy). It follows that a detailed measurement of the helical mean length at the melting temperature would give a reliable estimate of this important parameter. By studying the partial melting curves in detail, the temperature dependence of the bond free energies J_1 and J_2 can be confirmed.

Comparison with experimental results is ultimate for any theoretical model. For this particular venture, we have at worst obtained some qualitative agreement in most areas, not to mention some cases in which even quantitative agreement has been reached. We are able to produce a melting profile which is non-symmetric and consistent with the experimental observation that the profile tends to fall more steeply below the melting temperature than above it. Other calculations usually produce a symmetric curve.

Inspired by one of the expressions, a simpler experimental procedure for the estimation of percentage GC content is suggested. This procedure would eliminate the need to track down the entire melting curve, as in the conventional method.

Extraction of information leading to the disclosure of the base sequence is one of the primary aims of genetics students. This has been part of the reason for a detailed study on this transition process. In our present work we have reinforced the belief that a more thorough

investigation, particularly in the area of partial melting curves, may help to achieve this goal. We have already shown that during the transition process, regions of higher AT content tend to denature first. This is in contrast to the idea that denaturation takes place uniformly along the molecule. In addition, this immediately suggests that if the partial melting curves lie closely to each other, the base sequence of the DNA molecule must have a high degree of mixing between the AT and GC pairs. If the curves are separated widely from each other, the degree of mixing is correspondingly less.

Our discussion has been mainly centred around three parameters: two for the bonding free energies of AT and GC pairs (J_1 and J_2) and one for the stacking interaction (U). Most of the time we were restricted to a discussion utilizing the periodic boundary condition and the assumption that the total number of base pairs N is large enough so that effects due to the two ends of the macromolecule can be neglected. In view of the good agreement with experiment, this model is probably already not far off from the true situation. Nevertheless, further improvements and generalizations on this model can be made. As is evident from Eq.(3.6) to Eq.(3.8), the grand partition function can be expressed in terms of a product of N Q-matrices. With the removal of the periodic boundary condition, the calculation is more than just finding the trace of the matrix product; the off-diagonal elements ought to be evaluated as well. With the assistance of the computer, this could be done easily for any reasonable value of N. Similarly, the removal of the assumption that all pairwise stacking interactions are identical and generalization into six different U's (see Eq.(3.13)) would necessitate a numerical evaluation

TABLE 5.1

BLOCK	% LENGTH	% GC
1	4.98	47.5
2	3.39	61.0
3	1.99	43.2
4	1.99	56.3
5	1.99	44.0
6	1.79	56.0
7	3.39	44.4
8	2.19	55.0
9	5.38	45.4
10	2.39	54.3
11	6.97	45.9
12	3.19	53.6
13	8.96	46.67
14	3.39	53.0
15	4.18	52.3
16	9.56	48.3
17	5.38	51.5
18	9.36	49.1
19	6.57	50.5
20	7.97	50.0
21	4.98	47.5

Overall % GC = 49.47

of the grand partition function or $\bar{\lambda}$, the largest eigenvalue of Q, because the solution of Eq. (3.11) for $\bar{\lambda}$ is no longer as simple as Eq. (3.20). Apart from the sacrifice of an analytic solution, these modifications are quite straight forward.

Recently, there is experimental evidence that in DNA's of some organisms (e.g. calf thymus and some temperate phages) there exists individual segments each having roughly a random distribution of base pairs but with a different mean GC content (Shugalii et al, 1971; Yamagisha, 1970). The present theory can be generalized to cover this case in the following manner. If there are M such segments, the ith segment having $n_i N$ base pairs with a fractional AT content X_i, then the total fraction of disrupted base pairs, $\theta_t(T)$ is

$$\theta_t(T) = \sum_{i=1}^{M} n_i \theta^{(i)}(T)$$

where $\theta^{(i)}(T)$ is the fraction of disrupted base pairs in the ith segment, which can be evaluated in terms of X_i by the method described in previous chapters. After determining $\theta_t(T)$ the evaluation of other quantities is straight forward. A sample calculation is illustrated in Figs. 5.1 to 5.4. This calculation is based on the DNA of E. Coli (Yamagisha, 1970; Wartell and Montroll, 1972) which is taken to be made up of 21 unequal blocks each having a different GC content. Each block is assumed to have a random distribution of base pairs. The melting profile and other properties are compared with those of a hypothetical DNA having the same average GC content. It can be seen that these properties differ from each other mainly at the beginning and at the end of the transition. At

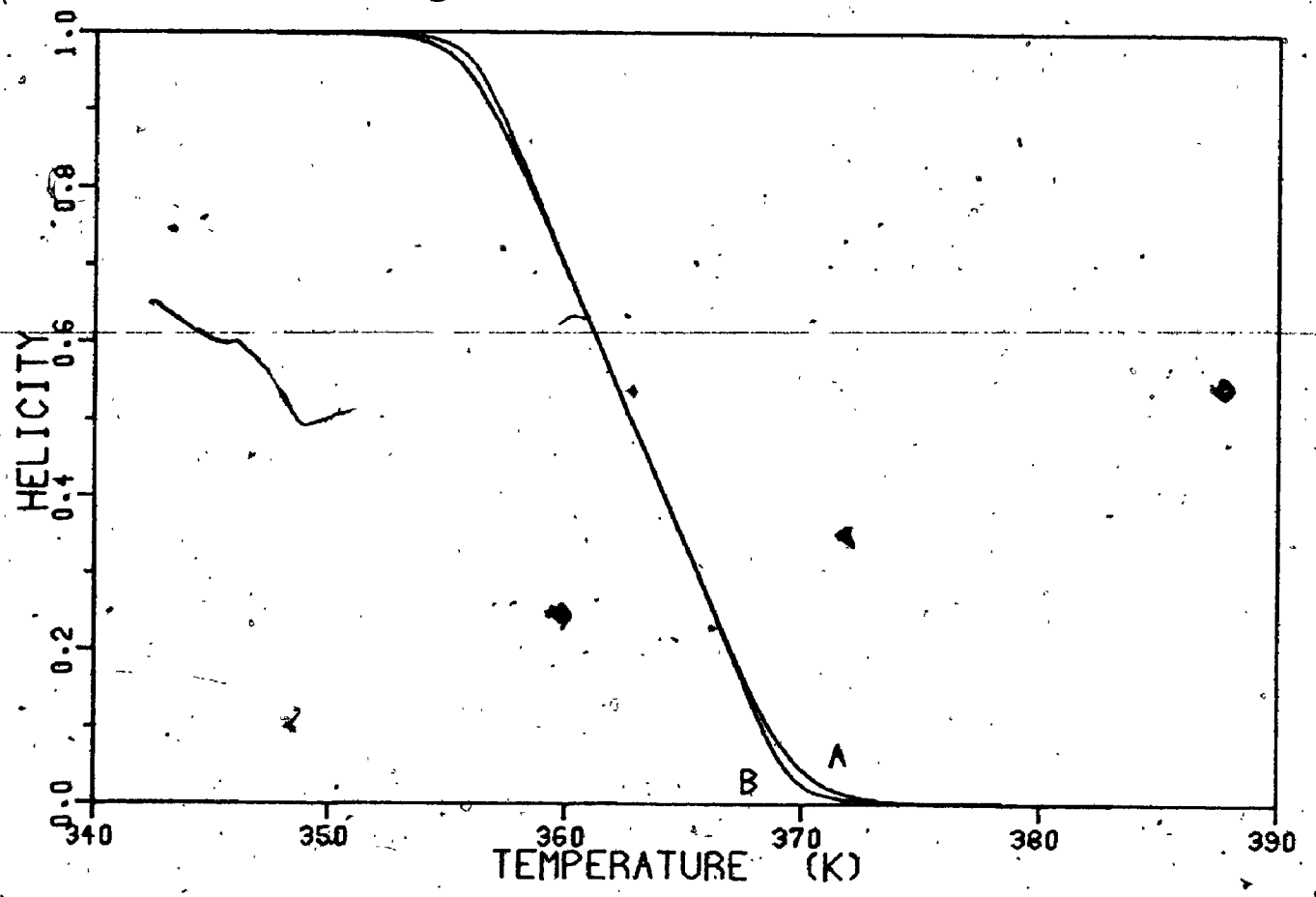


FIGURE 5.1 MELTING CURVES OF (A) E. COLI DNA OF 21 BLOCKS AND (B) DNA HAVING THE SAME GC CONTENT.

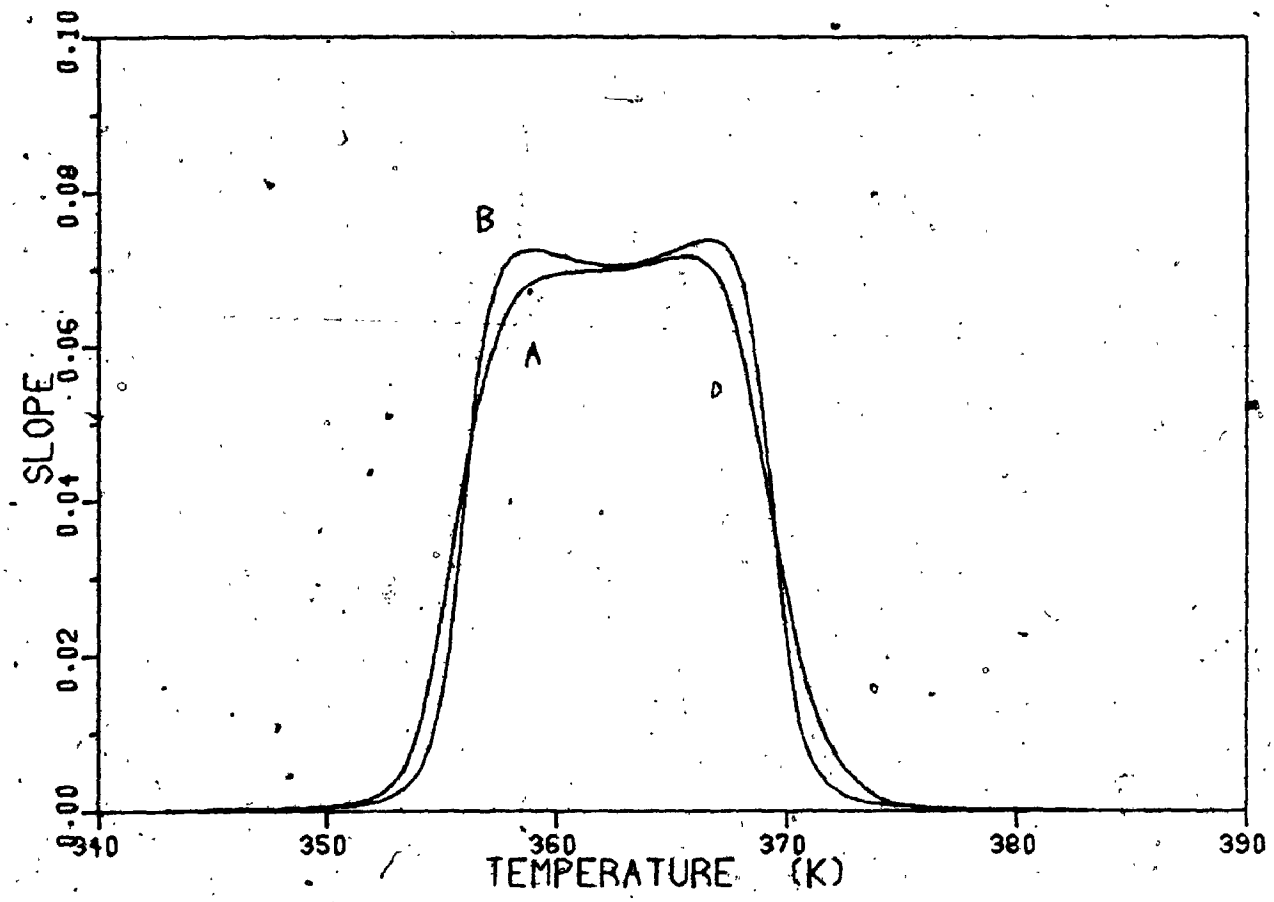


FIGURE 5.2 SLOPES OF MELTING CURVE OF (A) E. COLI OF 21 BLOCKS AND (B) DNA HAVING THE SAME GC CONTENT.

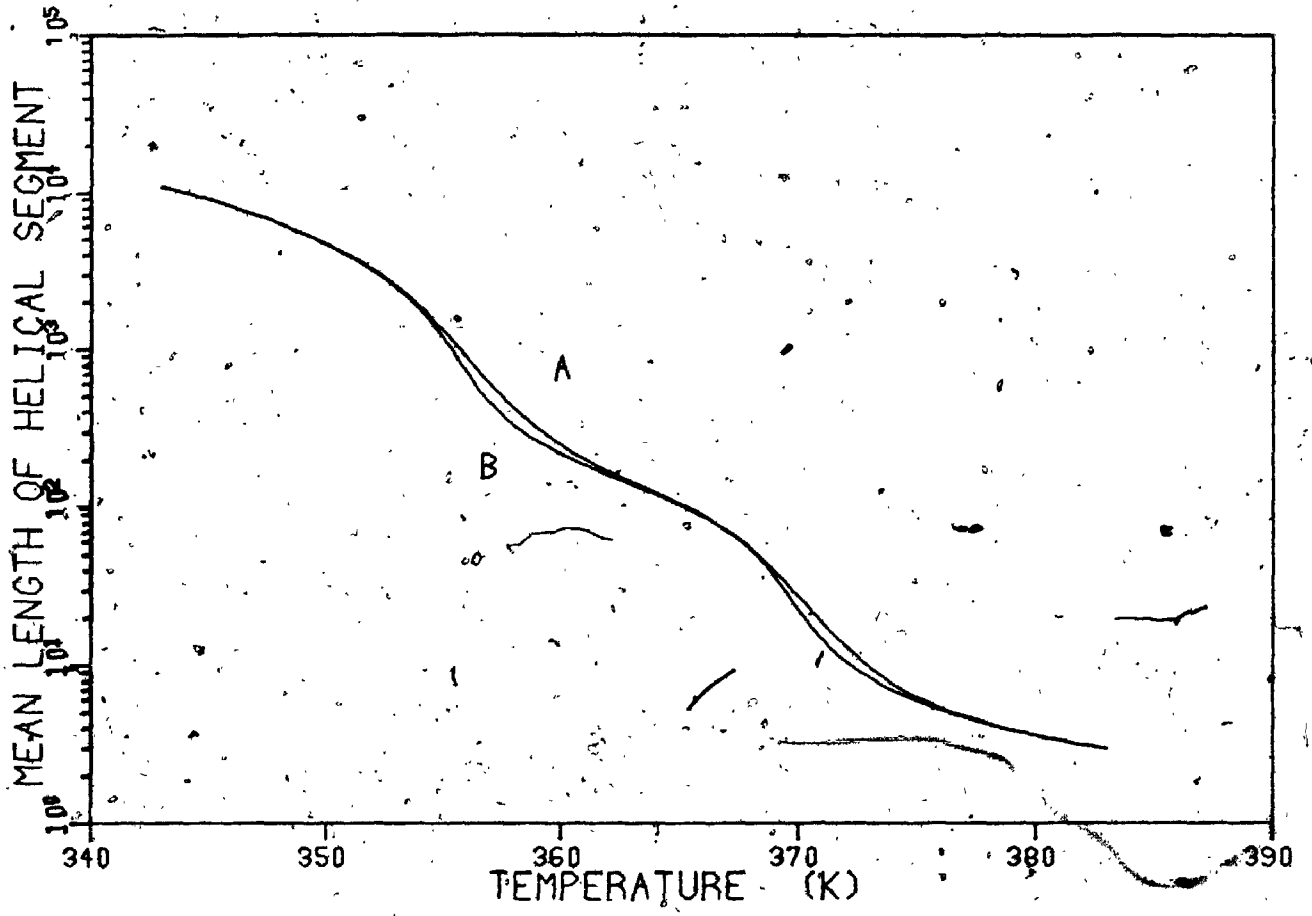


FIGURE 5.3 MEAN LENGTH OF HELICAL SEGMENTS OF (A) E. COLI DNA OF 21 BLOCKS (B) DNA HAVING THE SAME GC CONTENT.

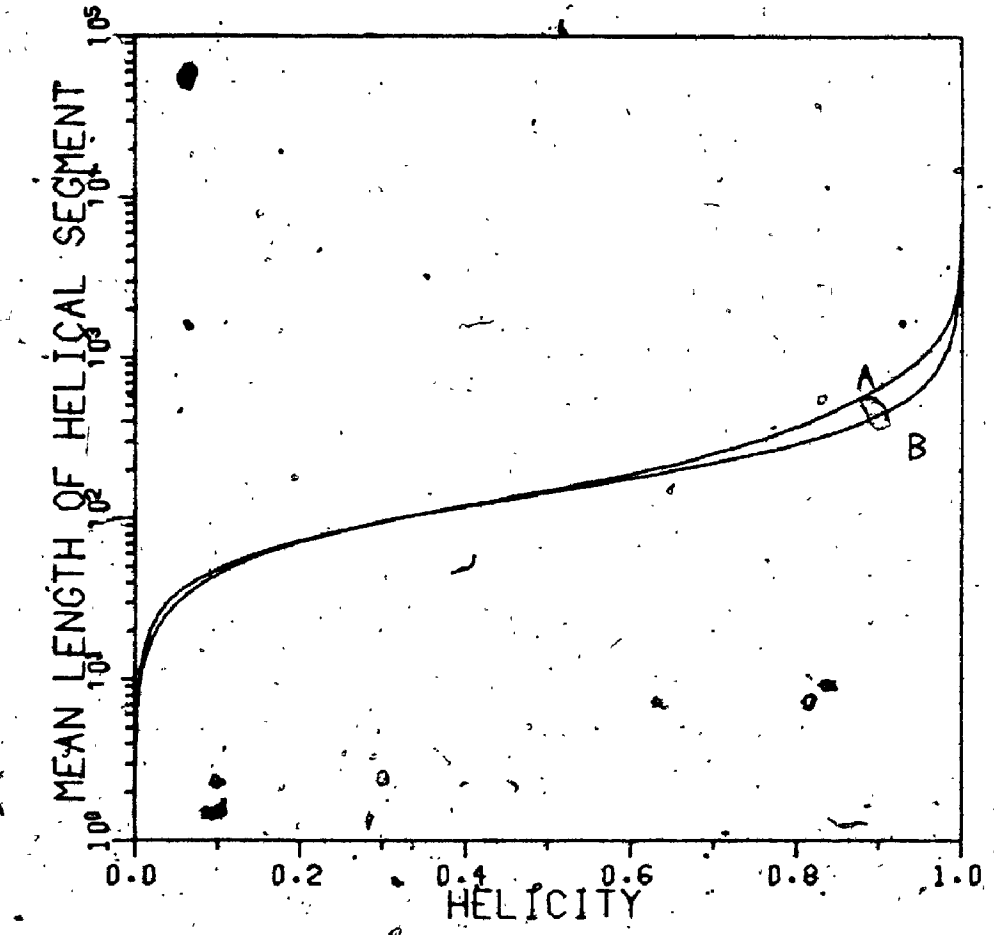


FIGURE 5.4 (A) DNA OF E. COLI (21 BLOCKS) AND (B) DNA HAVING THE SAME GC CONTENT

temperatures near the transition point T_m, they are not significantly different. This means that the block effects are not prominent in this region of temperature. However, block effects do make the transition sharper. Irregularities observed in the melting curves of some organisms may very well be due to this effect.

Recent progress has enabled us to establish the entire base sequence of some RNA molecules. It is hoped that in the foreseeable future the base sequence of some short DNA molecules may also be established. This, together with the discovery that a considerable population of DNA molecules do have repetitive patterns in their base sequences, emphasizes the approach proposed by Tong and Tong (1971) in studying the properties of DNA having a known sequence. Their approach is particular useful in studying these repetitive DNA's.

APPENDIX I

Instead of solving the quartic equation in θ or ϕ, we can also start with the following quartic equations in z, the fugazity ratio or the variable R:-

z-quartic:-

$$(1-X)^2(s_1^2+\sigma)z^4 + 2(1-X)[-s_1^2c_- + (1-X)s_1(s_1c_- + s_2) + \sigma c_-(1-2X)]z^3 + [(1-X)^2s_2^2 + X^2s_1^2 + \sigma(c_-^2 - 2X(1-X))]z^2 + 2X[-s_2^2c_- + Xs_2(s_1 + s_2c_-) - \sigma c_-(1-2X)]z + X^2(s_2^2+\sigma) = 0$$

R-quartic:-

$$(1-X)^2R^4 + 2X(1-X)s_1s_-R^3 + [-s_2^2 + \sigma(s_-^2 - 1) - 2Xs_-(c_1s_2 + \sigma s_-) - X^2s_-^2]R^2 - 2\sigma Xc_2s_-^2R + \sigma(s_2^2+\sigma)s^2 = 0$$

APPENDIX II

By definition,

$$\theta_1 = \left\langle \frac{\text{Number of broken AT pairs}}{\text{Number of AT pairs}} \right\rangle = \tfrac{1}{2}(1 - \phi_1)$$

$$\theta_2 = \left\langle \frac{\text{Number of broken GC pairs}}{\text{Number of GC pairs}} \right\rangle = \tfrac{1}{2}(1 - \phi_2)$$

$$\theta = \left\langle \frac{\text{Total number of broken pairs}}{\text{Total number of base pairs}} \right\rangle = \tfrac{1}{2}(1 - \phi)$$

It can be seen that

$$\left\langle \sum_j \tau_j \right\rangle = \frac{1}{Y}\left(\frac{\partial Y}{\partial J_1} + \frac{\partial Y}{\partial J_2}\right)$$

$$= \frac{N}{\lambda}\left(\frac{\partial \lambda}{\partial J_1} + \frac{\partial \lambda}{\partial J_2}\right)$$

$$\frac{\partial \lambda}{\partial J_1} = y\left[s_1 z + \frac{s_1 z(c_1 z + c_2) - (1 - \sigma)\, z s_-}{R}\right]$$

and

$$\frac{\partial \lambda}{\partial J_2} = y\left[s_2 + \frac{s_2(c_1 z + c_2) + (1 - \sigma)\, z s_-}{R}\right]$$

Now,

$$\theta = \frac{1}{2N} < N - \sum_j \tau_j >$$

$$= \frac{1}{2} \left(1 - \frac{1}{N} < \sum_j \tau_j >\right)$$

$$= \frac{1}{2} \left(1 - \frac{s_1 z + s_2}{R}\right)$$

$$\theta_1 = \frac{1}{2XN} < \sum_j \{g(\tau_j) - \tau_j^{AT}\} >$$

$$= \frac{1}{2XN} \{< \sum_j g(\tau_j) > - < \sum_j \tau_j^{AT} >\}$$

$$= \frac{1}{2XN} \left(XN - \frac{1}{Y} \frac{\partial Y}{\partial J_1}\right)$$

$$= \tfrac{1}{2}\left(1 - \frac{1}{X\lambda} \frac{\partial \lambda}{\partial J_1}\right)$$

$$= \tfrac{1}{2}\left[1 - \frac{y}{X\lambda} \left(s_1 z + \frac{s_1 z(c_1 z + c_2) - (1 - \sigma) z s_-}{R}\right)\right]$$

Where

$$g(\tau_j) = \begin{cases} 1 & \text{for an AT pair} \\ 0 & \text{for a GC pair} \end{cases}$$

and

$\{\tau_j^{AT}\}$ is the subsequence of AT pairs.

Similarly,

$$\theta_2 = \tfrac{1}{2}\left(1 - \frac{1}{(1 - X)\lambda}\frac{\partial \lambda_0}{\partial J_2}\right)$$

$$= \tfrac{1}{2}\left[1 \frac{y}{(1 - X)\lambda}\left(s_2 + \frac{s_2(c_1 z + c_2) + (1 - \sigma)\, z s_0}{R}\right)_0\right]$$

APPENDIX III

We wish to prove the validity of the following formulae:

$$G = \ell n(c_1 + s_1)\sqrt{\frac{1-\phi_1}{1+\phi_1}} = \ell n(c_2 + s_2)\sqrt{\frac{1-\phi_2}{1+\phi_2}} = \ell n\,\frac{\sigma^{\frac{1}{2}}\phi + \sqrt{1-(1-\sigma)\phi^2}}{1+\phi} \qquad \text{(III 1)}$$

Using Eq. (3.21), (3.22), (3.26), (3,27) and (3.28), we arrive at the following equations:-

$$[(c_1 - \phi_1 s_1) + (s_1 - c_1\phi_1)] = \sigma^{\frac{1}{2}}\phi\sqrt{1-\phi_1^2} \qquad \text{(III 2)}$$

$$(c_2 - \phi_2 s_2)\phi + (s_2 - c_2\phi_2) = \sigma^{\frac{1}{2}}\sqrt{1-\phi_2^2} \qquad \text{(III 3)}$$

Define

$$A_1 = \sqrt{(1-\phi_1^2)(1-\phi^2)} > 0$$

$$A_2 = \sqrt{(1-\phi_2^2)(1-\phi^2)} > 0$$

$$\alpha_1 = \tfrac{1}{2}\ell n\,\frac{(1-\phi_1)(1+\phi)}{(1+\phi_1)(1-\phi)}$$

$$\alpha_2 = \tfrac{1}{2}\ell n\,\frac{(1-\phi_2)(1+\phi)}{(1+\phi_2)(1-\phi)}$$

so that

$$\sinh(\alpha_1 + J_1) = \sigma^{\frac{1}{2}}\phi A_1^{-1}\sqrt{1-\phi_1^2} = \frac{\sigma^{\frac{1}{2}}\phi}{\sqrt{1-\phi}}$$

Applying the formula

$$\sinh^{-1}\left(\frac{x}{a}\right) = \ln\left(\frac{x + \sqrt{x^2 + a^2}}{a}\right)$$

we get

$$J_1 = \ln \frac{(1 + \phi_1)(1 - \phi)}{(1 - \phi_1)(1 + \phi)} \left[\sigma^{\frac{1}{2}} \phi A_1^{-1} \sqrt{1 - \phi_1^2} + \sqrt{1 + \sigma\phi^2 A_1^{-2}(1 - \phi_1^2)}\right]$$

$$= \ln \sqrt{\frac{1 + \phi_1}{1 - \phi_1}} \frac{\sigma^{\frac{1}{2}} \phi + \sqrt{1 - (1 - \sigma)\phi^2}}{1 + \phi}$$

similarly,

$$J_2 = \ln \sqrt{\frac{1 + \phi_2}{1 - \phi_2}} \frac{\sigma^{\frac{1}{2}} \phi + \sqrt{1 - (1 - \sigma)\phi^2}}{1 + \phi}$$

Therefore

$$\frac{\sigma^{\frac{1}{2}}\phi + \sqrt{1 - (1 - \sigma)\phi^2}}{1 + \phi} = e^{J_1} \sqrt{\frac{1 - \phi_1}{1 + \phi_1}} = e^{J_2} \sqrt{\frac{1 - \phi_2}{1 + \phi_2}}$$

Taking the natural logarithm of the above equation gives the desired result.

APPENDIX IV

h ≡ mean no. of base pairs in a helical segment

γ ≡ mean no. of base pain between initiation points of consecutive helical regions.

For a particular conformation $\{\tau_j\}$, the number of helical segments (which is the same as the number of coil segments) is

$$\sum_{j=1}^{N} \tfrac{1}{4}(1 - \tau_j \tau_{j+1}) = \frac{N}{4} - \tfrac{1}{4} \sum_{j=1}^{N} \tau_j \tau_{j+1}$$

$$\gamma = \frac{N}{\langle \frac{N}{4} - \frac{1}{4} \sum_{j=1}^{N} \tau_j \tau_{j+1} \rangle} = \frac{4}{1 + \frac{4\sigma}{\lambda} \frac{\partial \lambda}{\partial \sigma}}$$

$$= \frac{(1 - \theta)N}{N\left[\frac{1}{4} + \frac{\tau}{\lambda} \frac{\partial \lambda}{\partial \sigma}\right]} = \frac{2(1 + \theta)}{1 - \frac{\gamma}{\lambda} \frac{\partial \lambda}{\partial \gamma}}$$

$$= \frac{(1 + \phi)R(c_1 z + c_2 + R)}{2}$$

$$\sigma(z^2 + 2c_1 z + 1)$$

Using

$$\phi R = s_1 z + s_2$$

$$R^2 = (c_1 z + c_2)^2 - (1 + \sigma)(z^2 + 2c_1 z + 1)$$

we can eliminate z and R to obtain the following equation:

$$(1 - \phi)(1 + \phi)^2 \left[1 + \phi + \sigma(h - 1)(\phi(h + 1) - h + 1)\right] = 0$$

giving

$$h = \frac{1 + \sigma^{-\frac{1}{2}}\sqrt{1 - (1 - \sigma)\phi^2}}{\phi}$$

and

$$\gamma = \frac{2h}{1 + \phi} = \frac{2(1 + \sigma^{-\frac{1}{2}}\sqrt{1 - (1 - \sigma)\phi^2})}{1 - \phi}$$

APPENDIX V

GLOSSARY OF TERMS

Adenine. A purine, see purine.

Base pairs. Two bases (usually one pyrimidine and one purine), one from either component polynucleotide of the DNA molecule, are linked together by hydrogen bonds to form a base pair. This interaction is important for the stability of the DNA double helix. Usually the pyrimidine cytosine (C) is linked to the purine guanine (G) and the pyrimidine thymine (T) to the purine adenine (A). In this way the GC pair and the AT pair are said to be complementary.

Complementary bases. see base pairs.

Cooperativity. By positive cooperativity we mean that there is a tendency for base pairs to assume the same state as its neighbours. Negative cooperativity would mean that the units tend to align themselves in opposite states.

Denaturation. Certain classes of biological molecules (e.g. proteins, nucleic acids) in their natural or native state are long polymerised chains wound up in a distinctive helical manner. In this form they are biologically active. By mild treatment of heat, acid, alkali, urea solutions or detergents, they can be transformed from their native state to the denatured state, so that they are no longer biologically active. It is supposed that

denaturation consists mainly of unwinding of the helix, not breaking up of the polymer itself. The reverse process is renaturation. This transformation is also known as helix-coil transition or melting.

Cytosine. A pyrimidine, see pyrimidine.

Guanine. A purine, see purine.

Helicity. Fraction of intact bonds or base pairs in the helical state.

Heteropolynucleotide. A polynucleotide in which its purine or pyrimidine bases are different from one another.

Homopolynucleotide. A polynucleotide in which all its purine and pyrimidine bases are identical, e.g. poly dAT : poly dAT.

Melting curve. This is a curve showing the fraction of denatured base pairs as a function of temperature.

Nitrogenous base. see purine and pyrimidine.

Nucleic acids. One of a group of organic compounds present in the nucleic and cytoplasm of cells, as ribonucleic acid (RNA) and deoxyribonucleic acid (DNA).

Nucleotide. An organic compound consists of a purine or pyrimidine base linked to a sugar (ribose or deoxyribose) is a nucleoside. A nucleotide is a nucleoside phosphate: a nucleoside in which the phosphoric acid group is attached to a sugar.

Partial melting curves. This refers to a particular type of base pair in the DNA, say AT. The AT partial melting curve gives the fraction of AT base

pairs which has been denatured as a function of temperature.

Purine. A planar heterocyclic compound composed of a pyrimidine and an imidazole ring with two carbon atoms in common. Purines (and pyrimidines) are building units of nucleic acids, and they play the important role in the storage of genetic information in the process of cell division, reproduction and the transmission of hereditary factors.

Pyrimidine. It is a planar heterocyclic ring containing four carbon and two nitrogen atoms. Though it has not been found to occur naturally, some of its derivatives are widely distributed in nature. Accordingly, pyrimidines play a vital role in many biological processes.

Renaturation. see denaturation.

Stacking interaction. This refers to the base pair - base-pair interaction in the DNA molecule and is believed to be the main cause of cooperativity. In the case of the DNA molecule this stacking interaction causes the base pairs to align themselves together in a similar state (intact or disrupted).

Thymine. A pyrimidine. see pyrimidine.

REFERENCES

1. J. Bohacek, M. Kocur and T. Martinec, J. Gen. Microbiol., 46, 372 (1967).

2. J. Brahms, J.C. Maurizot and A.M. Michelson, J. Mol. Biol., 25, 451 (1967).

3. S.E. Bresler, Sov. Phys. Uspekhi, 12, No. 4, 543 (1970).

4. M. Boublik, L. Pivec, J. Sponar and Z. Sormova, Coll. Czech. Chem. Commun., 30, 2645 (1965).

5. M.F. Chiang, Ph. D. Thesis (1968).

6. S.S. Cohen and O. Penrose, J. Chem. Phys., 52, No. 10, 5018-5021 (1970).

7. D.M. Crothers and B.H. Zimm, J. Mol. Biol., 9, 1 (1964).

8. D.M. Crothers, N.R. Kallenbach and B.H. Zimm, J. Mol. Biol., 11, 802 (1965).

9. D.M. Crothers, Biopolymers, 6, 1391 (1968).

10. N. Davidson et al, Proc. Natl. Acad. Sci., 53, 111 (1965).

11. J. De Ley, J. Bact., 101, No. 3, 738-754 (1970).

12. J. De Ley and J. Schell, J. Gen. Microbiol., 33, 250 (1963).

13. W.F. Dove and N. Davidson, J. Mol. Biol., 5, 467 (1962).

14. S. Falkow and D. Cowie, J. Bact., 96, 777 (1968).

15. G. Felsenfeld and H.Z. Hirschman, J. Mol. Biol., 13, 407-427 (1965).

16. G. Felsenfeld and G. Sandeen, J. Mol. Biol., 5, 587-610 (1962).

17. T.R. Fink and D.M. Crothers, Biopolymers, 6, 863-871 (1968).

18. N.S. Goel and E.W. Montroll, Biopolymers, 6, 1391 (1968).

19. S.Z. Hirschman, M. Gillert, S. Falkow and G. Gelsenfeld, J. Mol. Biol., 28, 469, (1967).

20. K. Huang, Statistical Mechanics, Wiley, New York (1963).

21. R.B. Inman, J. Mol. Biol., 28, 103 (1967).

22. Yu. S. Lazurkin, M.D. Frank-Kamenetskii and E.N. Trifonov, Biopolymers, 9, 1253-1306 (1970).

23. G.W. Lehman, Statistical Mechanics, T.A. Bak, Ed., Benjamin, New York, 204 (1967).

24. G.W. Lehman and J.P. McTague, J. Chem. Phys. 49, 3170 (1968).

25. M. Leung and B.Y. Tong, to be published.

26. S. Lifson, Biopolymers, 1, 25-32 (1963).

27. S. Lifson and B.H. Zimm, Biopolymers, 15-23 (1963).

28. I.M. Lifshitz, Sov. Phys. JETP, 38, No. 3, 545 (1974).

29. M. Mendel, Annu. Rev. Microbiol., 23, 239-267 (1969).

30. J. Marmur and P. Doty, J. Mol. Biol., 5, 109 (1962).

31. E.W. Montroll and N.S. Goel, Biopolymers, 4, 855 (1966).

32. R. E. Owen, L.R. Hill and F.P. Lapage, Biopolymers, 7, 503 (1969).

33. M. Ozaki, M. Tahaka, Y. Kawai and E. Teramoto, Progr. Theor. Phys. Japan, 38, 10 (1967)

34. ibid, 18, 551 (1963).

35. D. Poland and H.A. Scheraga, Biopolymers, 7, 887-908 (1969).

36. D. Poland and H.A. Scheraga, Theory of Helix-Coil Transitions in Biopolymers, Academic Press (1970).

37. B. Pullman, Ed., Molecular Associations in Biology, Academic Press (1968).

38. H. Reiss, R.A. McQuarrie, J.P. McTague and E.R. Cohen, J. Chem. Phys., 44, 4567 (1966).

39. R.J. Seidler and M. Mendel, J. Bact., 106, No. 2, 608-614 (1971).

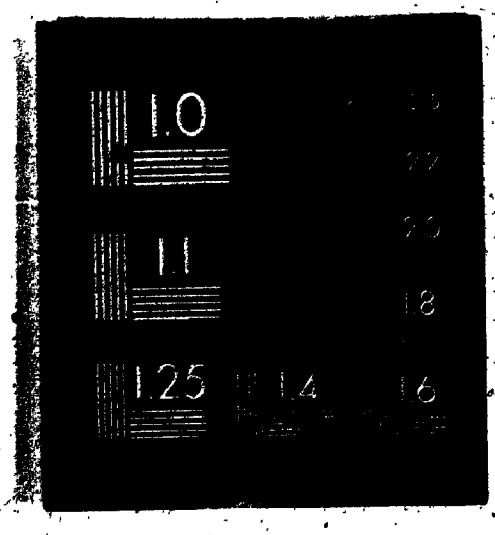
40. A.V. Shugalii, M.D. Frank-Kamenetskii and Yu. S. Lazurkin, Mol. Biol., 3, 103 (1969).

41. ibid, 5, 637 (1971).

42. R.F. Steiner and R.F. Beers Jr., Polynucleotides, Elsevier Publishing Co. (1961).

2 2

OF/DE



43. B.Y. Tong and S.Y. Tong, J. Chem. Phys., 54, No.3, 1317-1324 (1971).

44. A.A. Vedenov and A.M. Dykhne, Sov. Phys. JETP, 28, No.1, 187 (1969).

45. A.A. Vedenov, A.M. Dykhne and M.D. Frank-Kamenetskii, Sov. Phys. Uspekhi, 14, No.6, 715 (1972).

46. A.A. Vedenov, A.M. Dykhne, M.D. Frank-Kamenetskii and A.D. Frank-Kamenetskii, Mol. Biol., 1, 313 (1967).

47. J.D. Watson, Molecular Biology of the Gene, Benjamin, New York (1965).

48. J.D. Watson and F.H.C. Crick, Nature, 171, 737 (1953).

49. ibid, 171, 964 (1953).

50. G.H. Wannier, Statistical Physcis, John Wiley & Sons Inc. (1966).

51. R.M. Wartell and R.W. Montroll, Adv. Chem. Phys., 22, 129-203 (1972).

52. H. Yamagisha, J. Mol. Biol., 49, 603 (1970).

53. B.H. Zimm, J. Chem. Phys., 33, No.5,1349-1356 (1960).

54. B.H. Zimm and D.M. Crothers, Proc. Nat. Acad. Sci. (U.S.), 48, 905 (1962).